2020·广州
世界5G大会
WORLD 5G CONVENTION

THE NEXT-GENERATION PROTOCOL STACK

WHITE PAPER V7.0 B
2020.11

The Next-generation Protocol Stack

----RAN Architecture, Protocol Stack and Function

Version 3.0

Executive Summary

The Next-Generation Protocol Stack 2.0 White Paper mainly focuses on application scenarios of vertical industry to analyzes the new protocol stack architecture and optimization solutions for protocol layer functions. Also, the Protocol Stack White Paper 2.0 pays attention to the problems in the existing network to proposes the solutions such as new adaptive layers and functional units to explore new protocol stack architectures and functions.

On the basis of the previous white papers, the Next-Generation Protocol Stack 3.0 White Paper puts forward thinking and exploration on the protocol stack architecture and protocol layer functions. The main contents of this white paper are as follows:

(1) Firstly, from two perspectives of protocol stack architecture and protocol layer function, the deficiencies of existing protocol stack architecture and function in CU/DU cloudification, network slice and SBA in RAN are sorted out.;

(2) Based on the deficiencies of the existing protocol stack architecture and function imagine, the development trend of the next-generation protocol stack are analyzed from the perspective of service-based, component-based and intelligence-based RAN;

(3) Finally, the analysis of various scenarios and problems in the current network is given, also the solutions and suggestions for the next-generation protocol stack architecture and function enhancement are proposed.

This white paper aims to stimulate the academic and the industry to work on the future network protocol stack, form a consensus on the next-generation network protocol stack, as well as promote 6G innovation and ecological construction.

Table of content

1 Study Status

1.1 Weaknesses and Potential Enhancements in Architecture

1.1.1 CU and DU

Figure1.1 shows the architecture of 5G, In 5G, CN, Transport Network and RAN are defined separately connecting through standard interfaces, i.e. Ng, Xn, F1. This way is helpful for decoupling different network elements, then it also isolates the load balance of computing power.

With the deeply convergence of IT, DT and CT, the flexible schedule of computing power in cloud platform is a basic requirement. Obviously, the architecture of 5G should be deeply evolved in order to keep the pace.

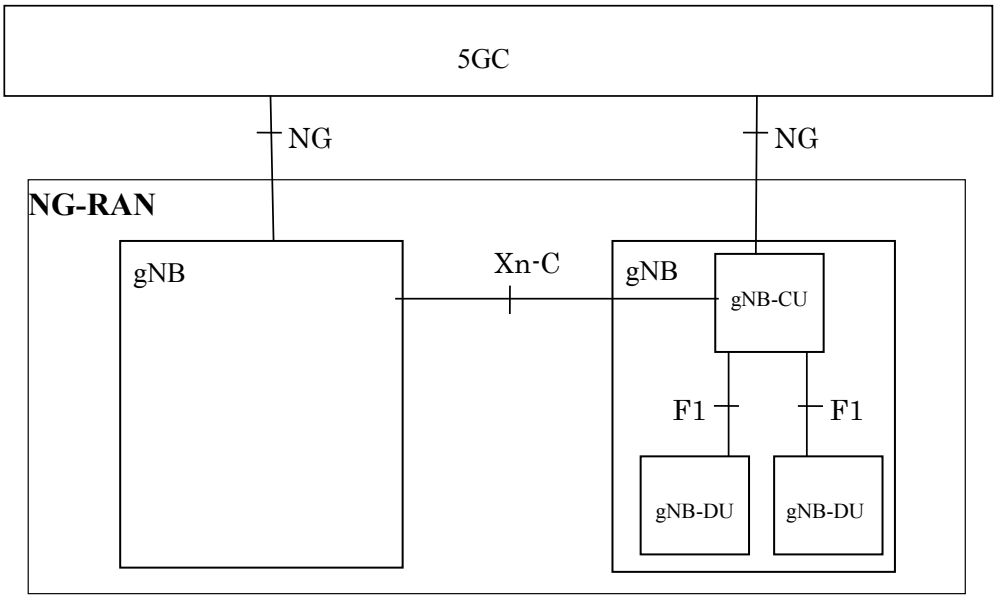

Figure 1.1 Overall Architecture

Figure1.2 and Figure 1.3 show the multiple connectivity scenarios in 5G which is a complex solution with redundant controlling.

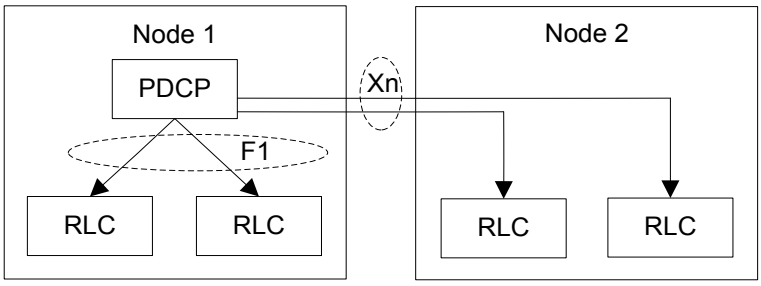

Figure 1.2 Scenario: DC+CA：2CA and 3DC

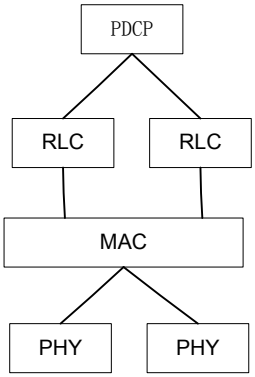

Figure 1.3 Overall Radio Link

The main problems are that the cell type definition is complicated (MCG/SCG/Pcell/Scell), and there are special split RBs, which introduce complicated bearer types and links. These problems not only increase the complexity of network management and signaling control, but also increase the handover delay and the handover failure probability. Also, due to the complex control of the F1 interface, the redundancy delay is high, and the protocol stack function is inefficient, which cannot adapt to the needs of future scenarios.

In addition, the current 5G network and protocol stack design are not flexible enough, and the openness is insufficient, making it difficult to adapt to the needs of multiple scenarios in the future. As the 5G core network, transmission network, and access network function definitions are independent of each other, it brings imbalance in control and functions, and the imbalance in centralization and distribution.

Each protocol stack layer runs independently, resulting in lack of cooperation between layers and insufficient flexibility. In addition, due to insufficient network openness, accurate opening of network capabilities cannot be achieved, which further makes it difficult to introduce new technologies such as big data and AI in the network.

1.1.2 NWDAF and slicing

Slicing is an important feature of 5G system and its main architecture advantages include shared infrastructure, resource security isolation, automated end-to-end resource and function configuration, on-demand customization, differentiated SLA, and end-to-end lifecycle management. For operators, it realizes end-to-end network resources sharing and flexible dynamic scheduling, dynamically optimizes network connections, effectively reduces the time for service activation, provides agile services, and reduces the cost of operators' network construction. For vertical industries and specific market segments, it has the ability to guarantee the quality of dynamic, differentiated services and provides different levels of security services. Furthermore, it enriches the slice-level flexible billing schemes and lays a foundation for future market cooperation and innovation. With the acceleration of 5G

commercialization, slicing has gradually become the focus of attention of vertical industries. However, in the process of its application, there are still many types of problems in design and application. This section will explain the challenges faced by the core network and wireless slicing.

- Technical challenges faced by core network slicing

(1) Comprehensive and intelligent slice management operation and maintenance[1-1]

Currently, the industry does not have a widely recognized network slicing management and orchestration system. Comprehensive, intelligent slicing management and orchestration should be an overall end-to-end system that can meet different service requirements while ensuring effective resource utilization and proper isolation. In order to achieve this goal, the system needs to efficiently and comprehensively manage resources based on the current state of the network slice, the state of underlying resource, and user's predicted needs. An efficient resource scheduling mechanism needs to be combined with optimal technologies to dynamically expand slice instance resources to meet user's changing needs without negatively affecting the performance of other slice instances. Specifically, considering user's ever-changing needs, distributed and dynamic cloud resource status, how to optimize deployment and dynamically adjust network slice resources is one of the most difficult challenges at present.

In addition, according to the needs of vertical industry customers, network slicing needs to refine the SLA requirements and decompose it into core network, wireless network, and transmission domains, and select the appropriate template for slice instantiation. The problems such as, how to effectively define SLA and how to ensure quality, are still lack of end-to-end monitoring and dynamic adjustment mechanisms to achieve closed-loop operation of network slicing.

(2) Lifecycle management automation

At the service level, life cycle management must be automated through closed-loop service assurance, fulfillment and orchestration functions covering all life cycle stages (as shown in Figure 1.4), including preparation phase, instantiation, configuration and activation phase, and operation-time phase and decommissioning phase. Two basic technology drivers include softening, such as the virtualization of network functions, and software-defined programmable network functions and infrastructure resources. The E2E service operation function interacts with the functions for the domain resource and the function management. The example domains include RAN, core networks, transport networks, NFV and MEC. In addition to orchestration, the closed-loop process for resource realization, resource assurance and network intelligence also includes building blocks within each management domain. At a more granular time and space level, domain-specific controllers include SDN controllers. The SDN controller can be programmed to effectively enforce policies and rules at the resource and function level.

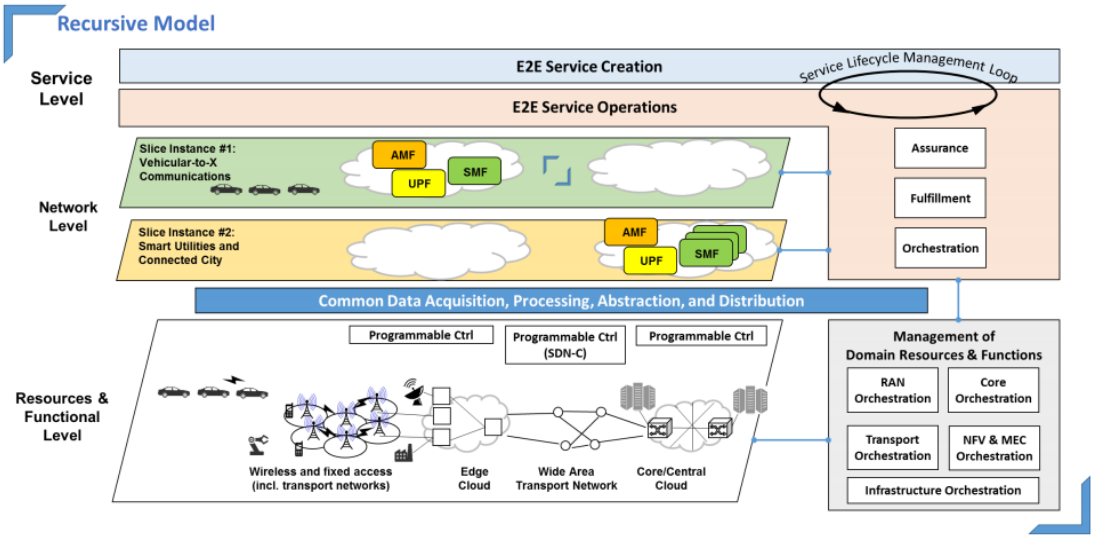

Figure 1.4 Automation of network slice lifecycle management

(3) Security challenge

Network slicing brings unprecedented security challenges, including security threats between slices and resource integration of network slices between domains. In addition, 3GPP has listed many security risks associated with 5G network slicing. First, the virtualized network function (VNF) of the network slice instance is deployed on a shared cloud-based infrastructure, so the security and performance isolation between different slices is essential. Since 100% physical isolation cannot be achieved, an attacker can flexibly consume the resources of another target slice by occupying a large amount of the capacity of one slice, thereby stopping the target slice from serving. In addition, according to the 5G architecture, certain control plane network elements, such as NSSF, are public to multiple slices, allowing an attacker to eavesdrop on the data of the target slice by illegally accessing the public function of another slice.

In addition, from UE's perspective, one UE can access multiple network slices at the same time, so UE may be exploited as a bridge to initiate security attacks from one slice to another. Judging from the current research and standardization of network slicing, relevant security mechanism has not yet formed a unified standard, and there is still a long way to go. This is also an important challenge for the commercialization of network slicing.

- Challenges faced by wireless network slicing

(1) Segmentation of the protocol stack

Traditional RAN resource scheduling is based on base stations. Air interface resources of the same base station are allocated to attached users under the action of a scheduler, but the scheduler cannot allocate resources in a detailed manner to user's needs. In the RAN side network slicing, in order to meet the specific air interface requirements of 5G scenarios, the radio access network needs to be segmented in order to be able to allocate resources more specific. According to the needs of different services, different protocols in the radio access network can be deployed to

different network elements (DU and CU) or the protocols that are not required by the service can be tailored, which can reduce data processing path distance. The segmentation of the wireless access network protocol stack is shown in Figure 1.5 [1-2]. Unicast services that do not require accessibility can tailor the radio link control (RLC) protocol. For Internet of Vehicles that require accessibility and delay, the RLC (Radio Link Control) protocol needs to be reserved, and the Packet Data Convergence Protocol (PDCP) should be deployed in the DU closer to the terminal to reduce delay. How to make the protocol stack reasonably and flexibly segmented is a technical issue worthy of study at present.

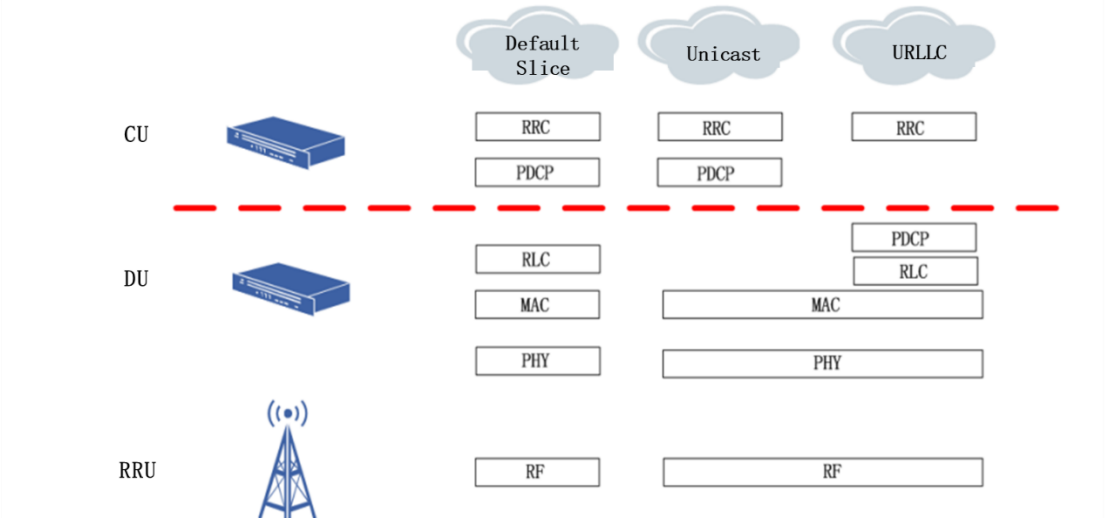

Figure 1.5 Segmentation of the wireless access network protocol stack

(2) Wireless network virtualization and resource scheduling

As a derivative of C-RAN, the 5G base station consists of one CU and multiple DUs. CU is more suitable for virtualization, but the function of DU is difficult to virtualize, because the function of DU relies heavily on dedicated hardware acceleration. Therefore, one of the biggest challenges lies in DU and physical channel virtualization. How to ensure effective slice isolation in beamforming is particularly worth studying. In addition, whether Multiple Radio Access Technologies (MRATs) can be virtualized through the same hardware or other dedicated hardware is particularly worth studying.

The results of a recent study [1-3] show that in order to achieve wireless resource scheduling based on slice granularity, new slice information, configuration descriptions and dedicated protocols need to be introduced to the layer 1/2/3 of the 5G wireless air interface. In order to achieve more effective slice isolation without affecting the multiple advantages of shared wireless resources, it is necessary to further explore virtualization and slice resource scheduling mechanisms. In addition, it is necessary to study whether the universal RRM applicable to all wireless slices or the dedicated RRM applicable to different slices separately, because it is difficult for a centralized RRM to manage different hardware components with different functional segments. The wireless resource scheduling scheme based on the granularity of network slicing is the focus of the next step.

(3) Isolation of RAN slicing

The random access method makes the isolation of virtual slices very complicated, however it has not been fully studied at present. Many aspects of isolation such as variable traffic, mobility, and variable channel bearing have not been fully investigated, so designing an effective slicing mechanism that ensures isolation is still a huge challenge.

(4) Decoupling of slicing and wireless technology

Different wireless access technologies have different air interfaces, spectrums, and protocols. Today, there is no unified method to deal with above factors. Therefore, it is a huge challenge to realize the resource allocation and isolation of wireless slice without considering the wireless technology.

(5) Real deployment issues

There are very few studies to deploy and test slicing schemes in real networks. In the wireless field, it is very important to perform a real deployment to evaluate the solution. For example, the multi-AP situation can bring new problems, such as the allocation and networking of control nodes, and it is necessary to consider which horizontal slice to choose.

(6) User mobility and interference issues

User mobility is a characteristic of wireless networks, and it will bring huge challenges to slicing. Mobility not only affects the connection capacity of nodes, but also causes obvious changes in the number of users carried by different nodes, which makes management very complicated. The wireless network has to deal with user mobility switching and ensure that the user's location will not affect the user's service quality.

(7) User access control

Different wireless access technologies have different access control methods, and the access control of users will also greatly affect the slicing performance of wireless resources.

(8) Wireless management function and configuration

Most wireless devices have complex management functions, including driver programming and low-level software. When multiple slices share a physical infrastructure, these functions should be used carefully to prevent slice conflicts from using commands.

(9) New technology compatibility

In order to meet the new demands of wireless networks, other new technologies are also emerging, such as extreme densification and offloading technology How to integrate slicing technology with other new technologies is also a challenge.

[1-1]李昊,张坦,王妮.5G网络切片技术综述与初期部署方案研究[J].通信技术,2020,53(05):1053-1062.

[1-2]古超智.5G端到端网络切片关键技术的探讨[J].广西通信技术,2019(03):10-13+17.

[1-3]Ferrus R.On 5G Radio Access Network Slicing:Radio Interface Protocol Features and Configuration[J].IEEE Commun. Mag.,2018

1.1.3 SBA in CN

Core network of 5G becomes softer than that of 4G by introducing Service-based Architecture (SBA). The observation is based on the independance of service-oriented functions in terms of scalability, extensibility, portability, customization , as well as the generality of service-based interfaces. From macroscopic perspective, SBA is based on the well-defined micro-service architecture and the RESTful architectural style in IT industry. It not only provides a complete vision, methodology, tools and concepts for 5G CN, but also provides a successful roadmap for the convergence of IT and CT.

Revisiting the history of SBA for the future, we can find the following lessons with the commercilization of 5G CN:

SBA was proposed upon the start of study item of next generation network architecture in 3GPP SA2, but not several years before the standardization process. Thus the important process of experimental and iterative cultivation, discussion, interaction and practice is missing. As a result, the standardization process is in a rush, some fundamental issues for the convergence of ICT is not addressed, and there's not enough time for iterative experiment and practice. Then some problems exist.

The Domain-Driven Design (DDD), as the guideline methodology of micro-service architecture, is not introduced in SBA.

The End-to-End SBA is not defined with the absense of SBA in RAN.

The balance of softness and efficiency is not well-defined, leading to less efficiency of service-based interface and the absence of service-based interface between control plane and user plane.

1.1.4 Absence of SBA in RAN

Partly due to the time scarcity of the introduction of SBA, the objective diversity of RAN functions on the trade-off of flexibility and efficiency, and the container-like virtualized platform for SBA is not suitable for all RAN functions, 5G RAN does not adopt service-based architecture. However, with the recognition progress of SBA, the requirement difference on platform and flexibility/efficiency of RAN functions becomes clear. Then the SBA in RAN becomes mature with the following considerations.

First, SBA in RAN, particularly the SBA of RAN control plane, should be considered as a part of the End-to-end SBA, together with the SBA in core network. Also the service capability for MEC and capability exposure should be supported.

Second, SBA in RAN should consider the balance of flexibility and efficiency.

Third, SBA in RAN should refer to Domain-Driven Design(DDD),the methodology of micro-service architecture and study the RESTful style deeply, in order to form a methodology suitable for radio access network.

1.1.5 Tight coupling of CU/DU and protocol functions

There are eight alternatives for CU/DU split architecture during the study phase. However, the mapping of SDAP/PDCP/RLC/MAC/PHY to CU/DU is static for any one of the eight alternatives. As the solution adopted by 3GPP, RRC/SDAP/PDCP is mapped to the CU and RLC/MAC/PHY is mapped to the DU. In another word, CU/DU and protocol stacks are tightly coupled. As vertical industries are likely to ask for the deployment flexibility, the tight coupling of CU/DU and protocol stacks restricts the flexibility of deployment.

For edge computing, usually CU/DU are collocated and integratedly deployed. For this case, the mapping of protocol stacks to CU/DU is not so important. For broadcast and multicast services, it is more efficient to deploy CU as a central node. For this case, it is suitable to map SDAP/PDCP to CU and RLC/MAC/PHY to DU. For video services with requirement of high reliability, it is appropriate to deploy SDAP/PDCP and the ARQ function of RLC to CU.

For next generation protocol, the functional partitioning of protocol stacks should be considered in a holistic view. The preliminary objective is to minimize the dependency among different protocols and decouple their services provided. Base on that, flexibility of mapping logical functions and protocol stacks to different entities should be supported.

1.2 Weaknesses and Potential Enhancements in Protocol Stack

In the Rel-16, 3GPP has studied the automation of 5G networks in SA. With the introduction of Network Data Analytics Function(NWDAF) in the CN side, various operation and maintenance data of the network are collected. According to the collected information, artificial intelligence technology is used to empower the mobile Communication network, optimize network parameters, improve network performance and user experience. In order to protect user privacy, most of the data collected are the network related information.

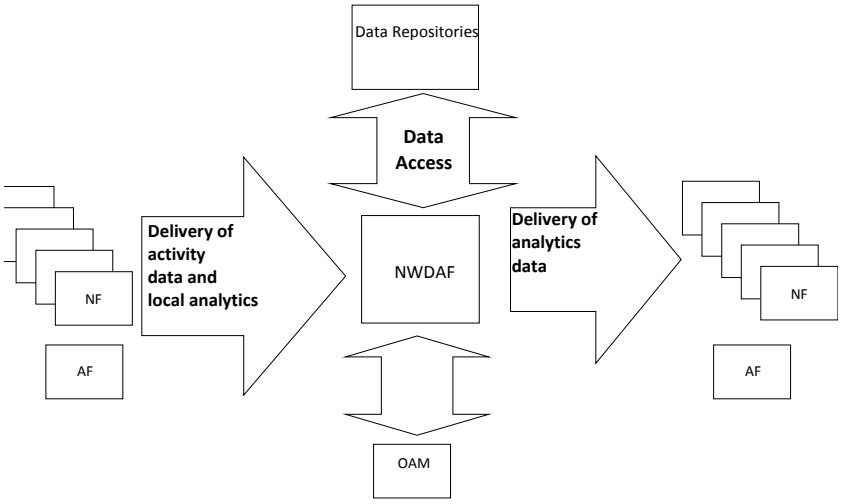

Figure 1.6 Network Automation for 5G

As we all know, the UE can collect a variety of user data, including the user's habits, daily movement track, etc. If these AI data can be used locally, it will not only improve user experience and network efficiency, but also avoid users' concerns about privacy leakage.

Currently, the communication protocol stack on the UE side does not support the interface with the AI module, that is, the AI module cannot learn the configuration information (such as network broadcast information) and link quality information from the network side through the communication protocol stack. In the opposite, the communication protocol stack cannot obtain user behavior information. Define related interfaces will be a potential direction for enhancement of the communication protocol stack.

2 Next Generation RAN Architecture

2.1 Service-based Architecture for RAN

The characteristics of softness for the next generation network is aligned with the cloudification trend in the network evolution. Here cloudification not only means the deployment in the cloud, but also means the network design based on Cloud-Native guideline, including portability via container, scalability, extensibility and customization based on mircro-service architecture, continuous integration and delivery with DevOps, and the organization structure supporting the network architecture with Conway’s Law.

With NFV or Platform as a Service (PaaS) in 5G, portability can be supported. The direction of future network is to further enhance the capability of scalability, extensibility, customization, and continuous integration. Specifically, next generation network should extend the service-based architecture for core network to End-to-End Service-based Architecture (E2E-SBA). For that purpose, the radio access network should be refactored with SBA to support SBA-RAN. The UE and radio interface should also be redesigned in certain aspects with the basic guideline of SBA (please note that it is not suitable to copy the service-based interface to the radio interface from the efficiency perspective).

With the requirement of extensibility, customization, and continuous integration, life-cycle management and life-cycle driven domain decoupling should be adopted in SBA design. Furthermore, functions with different attributes should also be separated. In this way, end-to-end service-based architecture should have the service fabric which is not related to specific access or core network functions. The service fabric can include service registration, discovery and communication functions, unstructured storage function and protocol stacks for service-based interface. Next, RAN functions can be partitioned according to service logic and life-cycles into PHY domain, PKT (packet) domain and HUB (control hub) domain. Each domain can be partitioned

according to functions attributes into enforcement plane, control plane, storage plane, intelligence plane, and OAM plane. Thus SBA-RAN have functions including PHY-E/PHY-C/PHY-S/PHY-I/PHY-O, PKT-E/PKT-C/PKT-S/PKT-I/PKT-O, HUB-C/HUB-S/HUB-I/HUB-O. With such design, functions with different attributes can have different implementation platforms and deployment locations. They can be also orchestrated scalably on-demand. Different domains with different service logic and life-cycles can be independantly driven by their evolved requirements for further enhancement and evolution. All functions can be continuously orchestrated, customized, configured, and integrated into a network system.

According to the partitioning principles and recommendations, data network (DN) such as IMS and MEC, and core network(CN) can be integrated with RAN more smoothly in the following aspects.

The implementation platform and deployment location of DN, CN and RAN can be merged together on-demand: DN, CN and RAN can share a common service fabric. In some cases like edge cloud, DN, part of CN functions and part of RAN functions can be deployed on the same platform and at the same place.

The service logic of DN, CN and RAN can be merged: DN, PKT-related function of CN(UPF, SMF, AF) can be integrated with the PKT functions of RAN. Even the HUB-like function of CN (AMF and PCF) can be integrated with the HUB functions of RAN (including access and mobility function).

End-to-End SBA and the merge of DN, CN and RAN are not only technical trend, but also pulled by new business mode of the future network. Different from the traditional To C-Native network, future network will be a To B-Native network and To C will be supported on the basis of To B, so as to provide service network driven by the demands from enterprises, governments, and social organizations. There based on the public platforms, private platforms operated or owned by enterprises, government and social organizations, or hybrid platforms, people, organizations, services, intelligent agents, and things can be fully connected and enabled.

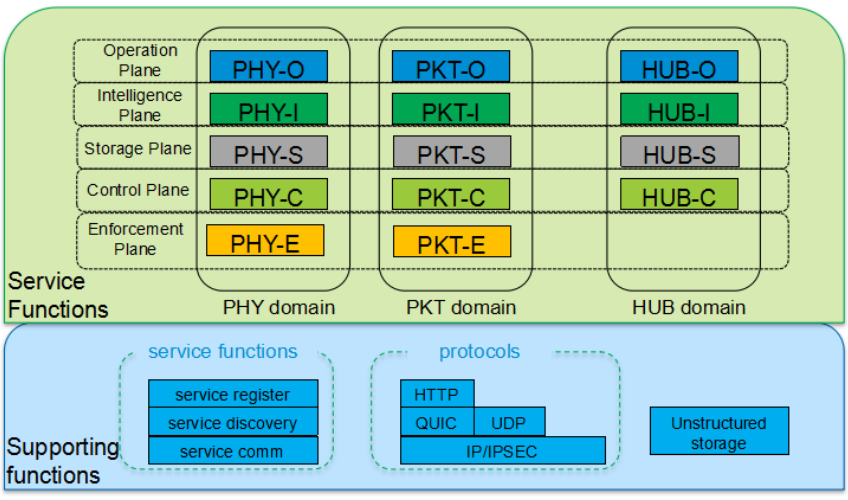

Figure 2.1 A holistic view of SBA-RAN

2.2 Component-based Forwarding Plane Architecture

For RAN, forwarding plane can forward user data and signaling. Thus forwarding plane can be considered as a more general plane than user plane.

Functions of forwarding plane should meet the basic requirement of security, efficiency, reliability, and QoS control. To support MEC, verticals, space-terrestrial integration, integrated access and backhaul, dual connectivity, carrier aggregation, URLLC, and interworking, forwarding plane should also meet the advanced requirement of generality, independence, compatibility, extensibility, portability and openness. For this purpose, component-based design is recommended for the forwarding plane. With the consideration of basic and advanced requirements, protocol stacks of forwarding plane can be partitioned into orchestratable modularized components, in order to achieve the balance of basic requirement (e.g. efficiency, reliability, etc.) and advanced requirement (e.g. compatibility, extensibility and openness).

Driven by the above requirements, PKT(packet)-related protocols are likely to evolve with component-based basic functions. Such functions can include basic packet encapsulation such as concatenation and segmentation, reliability control, security control, flow split and control, routing and mapping, address/name translation. Based on the components, packet service function chain (PKT-SFC) can be orchestrated and configured with on-demand customization.

2.3 Intelligence-based RAN Architecture

Intelligence is an important driven-force and enabler of next generation RAN. It impacts all RAN functions. Intelligence can be acted as an micro-scopic enabler within RAN functions, mesoscopic enabler by the side of RAN functions, and macroscopic enabler on the top of all RAN functions. Intelligence-based architecture will appear in different forms at different scales.

At the micro-scopic scale, intelligence can be embedded into existing functions, such as resource allocation function, so as to achieve better flexibility, adaptability, and higher performance with dynamic optimization. Regarding the architectural aspects, the embedding of intelligence may have impact on the interfaces among functions. On the one hand, interfaces will be used to collect the data required for intelligence. On the other hand, artificial intelligence may have revolutionary impact on the interaction ways among different functions, based on distributed intelligence architecture such as federated learning, or based on the revolution to basic functions such as physical layer by artificial intelligence.

At the mesoscopic scale, intelligence can be used as service-based functions. Based on the SBA-RAN architecture in the previous subsection, intelligent functions can provide services such as scenario analysis, problem diagnosis and strategy

assistance to other functions (such as control functions). They can also collaborate with each other across different domains, to achieve the integration of intelligent capability.

At the macroscopic scale, intelligence can be used as a holistic and enabling technology. Along with digital twin and intent-based networking, intelligence can be used to build a novel collaborative, visible, communicable interactive mode between people and network for total life-cycle management and operation.

One of the evolution directions of achieving intelligence-based RAN is combination of existing mobile communication systems and AI. And current implementation methods are mostly cloud-based centralized learning, that is, a large amount of training data is transmitted to the cloud, and the corresponding decision model is issued after deep learning. This brings delay problem, and the requirement for transmission bandwidth is very high and cannot meet the real-time requirements of the business. In future, with the substantial increase in computing resources on the edge access side and the evolving demand for the network of differentiated services, AI and access networks will continue to deepen integration.

First of all, intelligence is no longer a patch-style function after network deployment, only for performance optimization. Instead, it is considered at the beginning of the design. When facing different services, differentiated algorithm design, protocol stack design, and parameter design are all realization and adjustment through intelligence. It can be said that intelligence is closely integrated with the network element functions on the access side and achieves native AI. Secondly, in the future, the RAN side will be more integrated with edge computing to form an edge network. Intelligence will help realize edge network intelligence, including dynamic computing power allocation, providing ultra-low latency (jitter) and intensive computing resources for innovative services such as holography and autonomous driving.

3 Scenario and Protocol Stack Enhancement

In order to meet the requirement of the vertical industries, provide ubiquitous network services, and support more flexible network deployment, 3GPP has introduced supporting for some new features such as Time Sensitive Network (TSN), Non-Terrestrial Network (NTN) and Non-Public Network (NPN). The following is a brief introduction to the relevant features.

- **Time Sensitive Network**

At present, TSN technology has become a research hotspot in the Industrial Internet. Its delay guarantee and traffic reservation characteristics can guarantee the performance of industrial control services. At the same time, 5G networks have the advantages of flexible access and large connections, making the combination of TSN and mobile networks become possible. In order to support the existing TSN architecture, 5G was introduced as a TSN bridge and a fusion system that was not

perceived by the TSN network in 3GPP specification. At the same time, for clock synchronization and QoS mapping between the two systems, relevant technical solutions have been formulated. QoS guarantee is passed to the 5GS policy control function PCF network element through TSN AF. PCF is based on user's subscription and service flow requirements, and realizes the allocation of different QoS policies to service users of different levels. At the same time, the SMF and AMF network elements obtain the service QoS requirements issued by the PCF through the control plane signaling, and carry them to the RAN so that the wireless system can execute the corresponding scheduling strategy. Unlike ordinary service QoS, 5GC needs to transmit Time-Sensitive Communication Assistance Information (TSCAI)parameters to the RAN, such as burst arrival time, period, and flow direction, which are used for NG-RAN to reserve network resources for service flow. For clock synchronization, 5G clock domain, gNB can pass the clock to UE/DS-TT and UPF/NW-TT after obtaining the clock from the 5G GM (for example: GPS satellite) to achieve high-precision clock synchronization in 5GS. For TSN clock domain, UPF/NW-TT obtains TSN clock synchronization message from TSN GM, and UPF/NW-TT also needs to forward the clock synchronization message to UE/DS-TT through a specific QoS flow on the user plane, so as to realize the clock synchronization between UE/DS-TT and TSN GM.

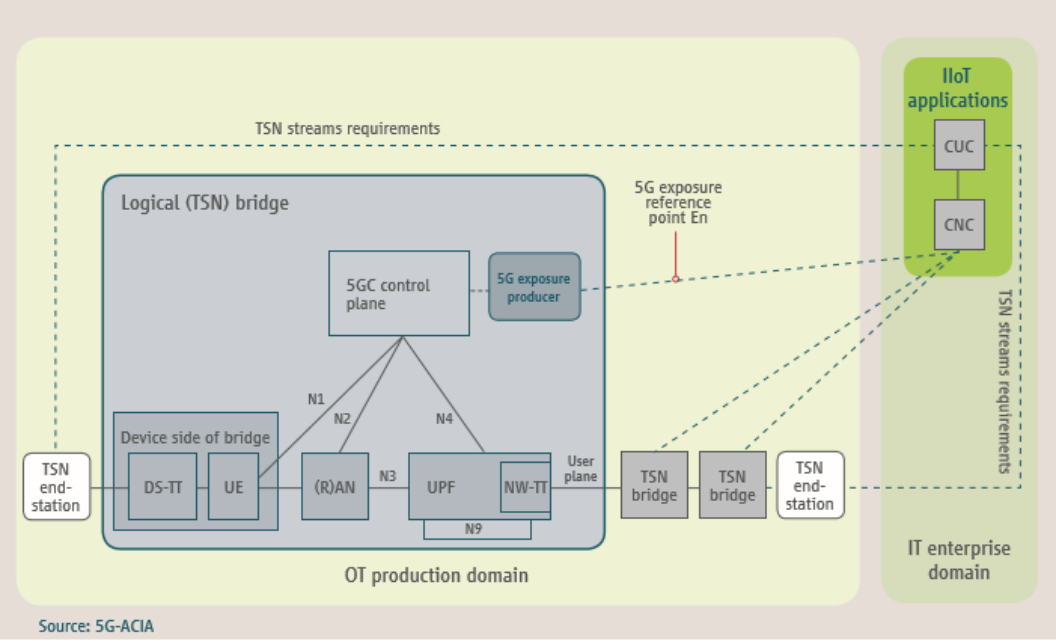

Figure 3.1 5G integrated into TSN

- **Low latency**

In TSN, the transmission time of most applications has fixed period. For periodic service flows, the semi-static resource can be configured in order to reduce the signaling overhead and extra latency of dynamic scheduling. In the 5G Rel-16, the configurable period of semi-static resource is extended to adapt to various periods of TSN traffic flows. Besides, a cell is enhanced to activate multiple semi-static resource configurations simultaneously to enable the transmission of parallel TSN traffic flows.

- **High reliability**

The reliability requirement of TSN service flow in industrial applications is up to 99.999999%. To support highly reliable transmission, 5G introduced PDCP duplication function, i.e. PDCP layer makes several copies for each data packet to be transmitted and the copies are delivery via different paths, i.e. cells or cell groups. The PDCP duplication can be configured in the same cell group, i.e. CA duplication, or in different cell groups, i.e. DC duplication.

The PDCP duplication function was introduced in Rel-15 with two transmission paths, and further enhanced to support four transmission paths in Rel-16 to support TSN service flow.

- **High efficiency**

The control information generated by industrial applications usually have a small size (e.g. a minimum of 20 bytes). Thus, the size of Ethernet frame header accounts for a relatively high proportion of the whole packet. 5G Rel-16 introduces the Ethernet Header Compression (EHC) function at the PDCP layer to reduce the overhead. Similar to the Robust Header Compression(ROHC) mechanism, EHC is based on removing the redundant information among data packets in a data stream. Different from ROHC, the EHC mechanism only compress the fields whose contents are static among packets. As a result, the EHC function is simple and easy to implement.

- **Space-ground integrated communication**

In practice, the deployment of 5G terrestrial wireless access networks has many restrictions. For example, it is difficult to deploy 5G networks in areas such as isolated islands and deserts; it is difficult to achieve cost-effective 5G network deployment in remote areas such as rural areas. In these cases, satellite communications can be considered as a supplement to the terrestrial wireless access network to provide communication services.

The 3GPP began to study the feasibility of integrating satellites into 5G from Rel-14, and conducted research on the mechanisms in the successive Rel-15 and Rel-16[3-1]. In Rel-17, 3GPP will start standardization work related to satellite communication mechanism enabled by 5G.

Reference：

[3-1] TR 38.821 Solutions for NR to support non-terrestrial networks (NTN)

➢ **Working modes**

A Non-Terrestrial Network (NTN) can provide services to users based on two modes: transparent payload and regenerative payload. The specific scenarios are shown in the following figures.

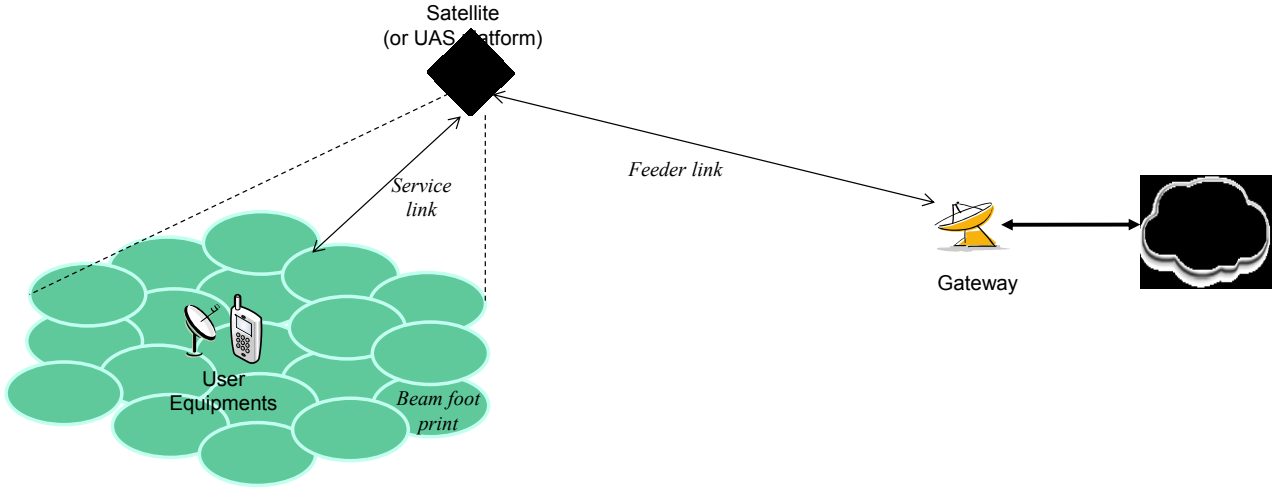

Figure 3.2 NTN scenario based on transparent payload

For NTN based on transparent payload, the satellite only equips with radio frequency filtering, frequency conversion and amplification. Thus, the satellite doesn't apply 5G access protocol layer.

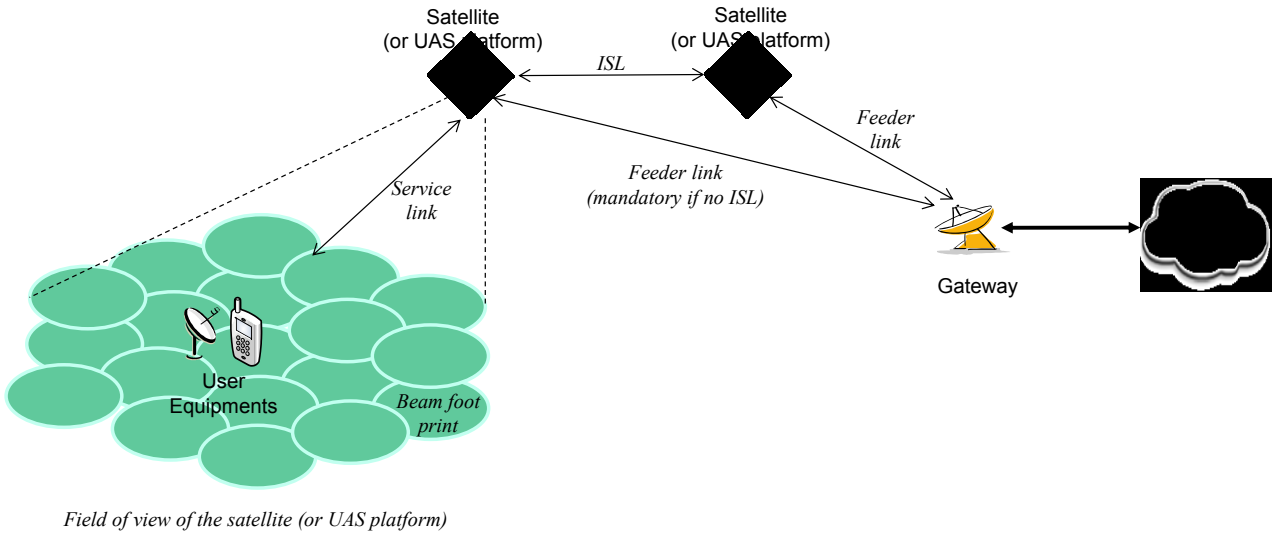

Figure 3.3 NTN scenario based on regenerative payload

For NTN based on regenerative payload, the satellite also equips with complete 5G access protocol layer. In the Rel-17, 3GPP will only focus on standardization of NTN based on transparent payload.

- **Management of Mobility**

In Rel-17, two scenarios will be supported: Low Earth Orbit satellite (LEO) scenario (with altitude of 600km and 1200km) and Geostationary Earth Orbit satellite (GEO) scenario. For the LEO scenario, since the coverage of satellite moves along with the movement of the satellite, moving cells are generated on the ground by the satellite. For the GEO scenario, since the satellite and the earth remain relatively stationary, fixed cell coverage can be achieved on the ground.

For the LEO scenario, to facilitate tracking area management, it was agreed to adopt a fixed tracking area, that is, the tracking area remains unchanged in terms of geographic location although the cells are moving. As a result, the network needs to maintain the fast-changing mapping relationship between the tracking area and the moving cells. Currently, two solutions, i.e. the soft and hard tracking area update mechanisms, have been proposed. The final solution still needs further discussion.

For mobility management of the Idle/Inactive UE, RAN-related enhancements include satellite ephemeris based cell selection/reselection; for mobility management of the Connected UE, RAN-related enhancements include RACH-less handover, conditional handover, etc.

- **Propagation delay**

Unlike terrestrial communication systems, the propagation delay of the NTN system is very large. Due to the difference of some factors such as payload types and satellite's altitude, the propagation delays vary greatly and the maximum one way(i.e. UE->satellite->gateway) propagation delay can reach 270ms. In order to adapt to the large propagation delay, some RAN functions need to be enhanced, including: uplink timing advance acquisition process, random access process, DRX function, HARQ function, uplink scheduling function, etc.

- **Feeder link switch**

Feeder link is used to connect satellite and the NTN gateway on the ground. For LEO scenarios, the satellite needs to switch its feeder link connection among different gateways while moving. There are two schemes for feeder link switch: soft switch and hard switch. For soft switch, the satellite keeps connection with the source and target gateway at the same time during the switching. For hard switch, the satellite firstly disconnects with the source gateway and then establishes the connection with the target gateway.

- **Non-Public Network**

Non-Public Network (NPN) can provide private entities, such as enterprises and schools, with exclusive network services to meet the requirement for low-latency and high-reliability.

3GPP introduced the NPN in Rel-16, which can be deployed in the following ways:

- Stand-alone NPN (SNPN): which does not rely on a PLMN. It can be operated by an SNPN operator or managed by a traditional operator.
- Public Network Integrated NPN(PNI-NPN): which rely on a PLMN and is operated by a traditional operator. The PNI-NPN can exist as a feature of traditional networks. The concept of Close Access Group (CAG) is introduced to realize logical resource separation for CAG and Non-CAG member UEs.

The NPN feature introduce no impact on the architecture of 5G protocol. NPN related standardization work mainly focuses on access control and mobility management. In order to ensure the dedicated resources for NPN UEs, the access control function is enhanced to restrict non-NPN UEs from attempting to access the NPN network. To avoid NPN UEs occupying the resources of non-NPN UEs, the access control function is enhanced to restrict NPN UEs from attempting to access the PLMN network.

To facilitate the UEs to identify whether the cell can provide NPN access, a NPN capable cell needs to broadcast NPN related ID. An SNPN is indicated by a PLMN ID and Network Identifier (NID), while a PNI-NPN is identified with a PLMN ID and CAG ID.

- **Standalone Non-Public Network**

For each SNPN service, the corresponding Subscription Permanent Identifier (SUPI) and credential are configured to the UE which has subscribed to the service. A SNPN-capable UE can work in PLMN access mode or SNPN access mode. When operating in the PLMN access mode, the UE behaves same as a traditional UE, i.e. it selects a cell that can provide PLMN access. When operating in SNPN access mode, the UE selects a cell that can provide SNPN access based on the SNPN subscription information. In network selection, the UE operating in SNPN access mode performs the SNPN selection process instead of the PLMN selection process [3-2].

In Rel-16, the interoperation of SNPN and PLMN is also standardized. Specifically, a UE operating in SNPN access mode that has successfully registered to SNPN can also perform PLMN registration through the SNPN user plane to obtain PLMN services, i.e. the interactive signaling and data between UE and the PLMN are transmitted through the SNPN user plane channel. Correspondingly, in order to obtain the SNPN service, a UE not operating in SNPN access mode that has successfully registered the PLMN can also perform SNPN registration through the PLMN user plane.

Reference:

[3-2] TS 23.122 NAS Functions related to Mobile Station (MS) in idle mode

- **Public Network Integrated Non-Public Network**

For UEs that have subscribed to the CAG service, a CAG list information (i.e. the UE is allowed to access) and the indication information of whether the UE can only access the network through the CAG cell (CAG only indication) are configured to UE per PLMN. These CAG related configuration information will be used as part of the UE's mobility restriction information and provided by the core network to the terminal's serving base station. The serving base station can manage the subsequent mobility of the UE based on these CAG related configuration information.

- **Emergency service**

In 5G Rel-16, 3GPP has standardized emergency services in NPN scenarios. The network may allow UE without NPN access capabilities (e.g. Rel-15 UE) to initiate emergency services from a PNI-NPN cell. However, SNPN cells cannot provide emergency services yet. The standardized work for providing emergency services in SNPN may to be included in the Rel-17.

3.1 NPN Enhancement

NPN has two methods for implementation: SNPN, which does not rely on PLMN, and PNI-NPN, which relies on PLMN. The protocol stack enhancement of NPN mainly focuses on radio interface and network interface.

1) Radio Interface

Firstly, NPN IDs need to be broadcasted in the SIB to make sure NPN UE could camp on the NPN Cells. To facilitate the manual selection for UE, HRNN (Human-Readable Network Name) can be involved mapping to NPN ID. In 3GPP Release 16, a cell can broadcast 12 NPN IDs at most. In the future, with the development of NPN, a cell should support more NPN IDs for resource sharing. In order to save UE power and speed up cell searching, PCI of neighbor cells supported the NPN can also be added into the SIB of PNI-NPN. When UE moves to the adjacent area, the scope for cells to access is narrowed down.

Secondly, the principle for cell selection/reselection is different between SNPN and PNI-NPN. SNPN follows the NR-U strategy. When the highest ranked cell is not the suitable cell, it will continue to search for the cell with the second rank at the same frequency. Only when the second ranked cell is not the suitable cell either, it will give up this frequency for cell search. PNI-NPN follows the traditional PLMN strategy. When the highest ranked cell is not the suitable cell, it will search suitable cells in other frequencies directly.

Thirdly, NPN should support finer access control in the future. Current specification only supports to make access control at PLMN level, which can not meet the differentiated CAG access control demand under the same PLMN. Thus, the future access control of NPN will be finer and more flexible.

As mentioned above, the main procedure for NPN-capable UE in radio interface could be shown as follow:

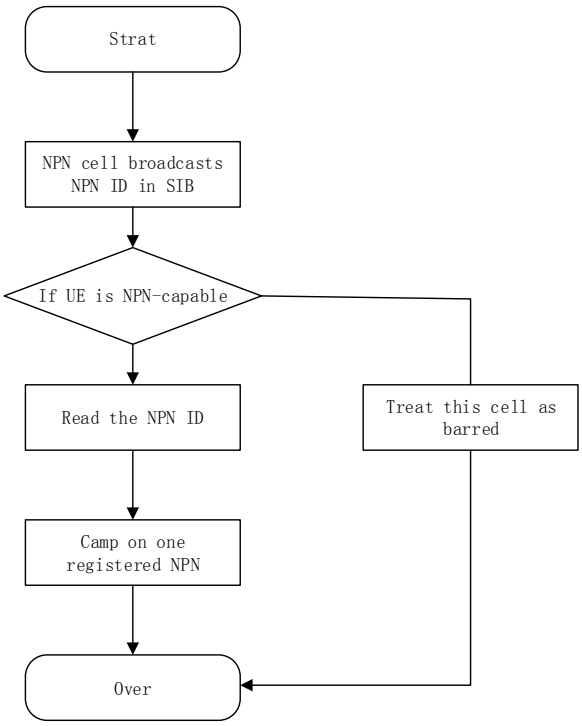

Figure 3.4 Procedure of NPN-capable UEs in radio interface

2) Network Interface

Except for radio interface, the network interfaces including Xn、F1 and NG also needs to be enhanced.

In NR system, gNBs could utilize Xn interface to interact supported NPN IDs with each other. Accordingly, if the configuration of NPN changes, it needs to be updated through Xn interface. In order to save the paging resources of the air interface, the NPN ID registered by UE is added to the paging signaling of Xn interface along with the indication of whether the current UE can only access to the NPN. Paging messages are only broadcasted among the UE registered NPN cells.

During handover procedure, the NPN related information is added to the handover request sent by the source gNB. For the SNPN handover, the serving NID is sent to the target gNB. If the target gNB does not support this SNPN type, the handover request is rejected. Otherwise, the handover request is allowed. For PNI-NPN handover, the list of allowed PNI-NPNs is included in the handover request message. It is up to the target gNB to respond whether the handover is allowed.

The enhancement of F1 and NG interfaces are basically the same as Xn interface. Both of them include the interaction of NPN information, such as paging and handover procedures.

3.2 Intelligent Network Enhancement in RAN

One operator may deploy and manage multiple radio access networks, such as NR, LTE, NB-IoT, WLAN and so on. Also, radio access technology becomes more and more complex. It is full of challenges in terms of network management and network optimization. 3GPP introduced self-configuration in Release 8 and self-optimization in Release 9 for LTE to support deployment of the system and performance optimization, which may be the earliest features related to intelligent network specified and used in 3GPP network. These features help operators reduce the CAPEX/OPEX and improve the user experience. For NR, 3GPP introduced some self-configuration features in Release 15, e.g. automatic neighbor relations, and introduced self-optimization features, such as mobility robustness optimization, mobility load balancing, RACH optimization, in Release 16, which also applies for ENDC. Besides the traditional self-configuration and self-optimization features, NR also introduced energy saving. The NG-RAN node owning a capacity booster cell or OAM can autonomously decide to switch-off such cell to lower energy consumption, based on cell load information, consistently with configured information.

NR uses the framework and solutions of self-configuration and self-optimization in LTE as baseline, and then takes the NR new architectures and features into account, e.g., MR-DC, CU-DU split architecture, beam. In august 2020, the standardized work of data collection for self-organizing network enhancement for NR and MRDC began in Release 17. This work will introduce other traditional self-configuration or self-optimization features for NR which hasn't been standardized, such as PCI selection, coverage and coverage optimization. On the other hand, it will study how to support network self-optimization of new features introduced in Release 16, such as 2-step RA, dual active protocol stack, conditional handover. However, due to the introduction of beam based antenna structures, the set of configurable antenna and RF parameters are multi-dimensional, which makes it very complex to find the mapping between network configurations with target coverage and capacity performance. Support of multiple system parameters and various network architectures in NR also increase the complexity of self-optimization algorithms. Some kind of machine learning techniques may be used to help the self-optimization function to analysis potential network problems based on the data collection in the radio access network.

On the other hand, 3GPP SA group started the study of concept, requirement and solutions for network automation in 2019, which considers a systematic way to evaluate how intelligent a network is. The study describes network autonomy as the telecom system capability which is able to be governed by itself with minimal to no human intervention, and defines network autonomy level as the level of application of autonomy capabilities in the network management workflow. The workflow consisted of one or more management tasks is shown as figure 3.5[3-3], which describes the necessary steps to achieve certain management purposes. The framework of network autonomy level classification based on autonomy capabilities of the tasks in the workflow is consisted of 6 levels, i.e. level 0 'manual operating network', level 1

'assisted operating network', level 2 'preliminary autonomous network', level 3 'intermediate autonomous network', level 4 'advanced autonomous network', level 5 'full autonomous network'. The higher autonomous network level may request more sufficient network or service data to improve the capability on awareness and request faster configurations to improve the capability on execution, therefore lead to the enhancement of data collection and configuration related interfaces. For 3GPP RAN, the features related to network autonomy, i.e. SON/MDT, are involved with 3 management tasks including network awareness, decision and execution, according to the work flow for network autonomy. To support network autonomy for network awareness, measurements enhancement can be introduced to MDT features to enable automatic network data collection. To support network autonomy for decision and execution, automatic UE reporting, inter-node information exchange, interface and network configuration enhancement can be introduced to SON features. Different autonomy capabilities of the SON/MDT features may lead to different capabilities of autonomy on awareness, decision and execution, which may lead to different autonomous network levels. Nevertheless, the network automation level of NR network is limited as it lacks systematic modeling and consideration for network automation at the beginning of NR design.

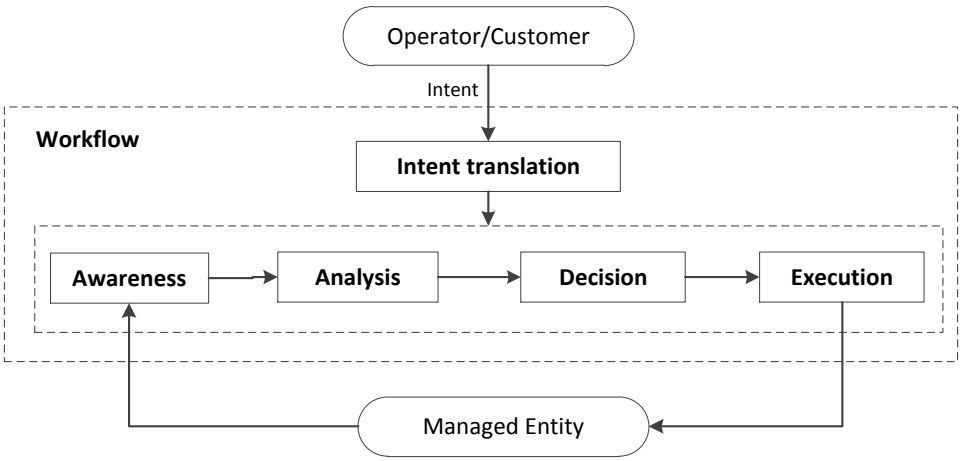

Figure 3.5 General workflow for network autonomy

Network automation can be considered as an important aspect in the 6G system design. For 6G, digital twin technologies may be used to build a virtual network which can reflect the status of the physical network, identify the need of adjusting network configuration and network performance optimization, provide solutions to optimize the network configuration and performance, and even predict the future state of the physical network. A system method involved with digital twin, big data analysis, and artificial intelligence technologies, will make it possible to achieve network automation.

Reference:

[3-3] 3GPP TR 28.810: 'Study on concept, requirements and solutions for levels of autonomous network'.

3.3 Time Sensitive Network enhancement

In the future, with the continuous diversification and complexity of industrial scenes and related services, they will be more sensitive to delay and service flow requirements, and services need to be guaranteed in every link of the network. In the subsequent evolution of the mobile network, the service support of the TSN network needs to be fully considered. Firstly, at the network architecture level, the mobile network and the TSN network should be deeply integrated, including the use of the TSN protocol at the fronthaul and backhaul network levels of the bearer network. Corresponding control and management network elements can be added to the mobile network system, responsible for the coordination of TSN and 5G control and management functions. Combined with MEC, local processing is performed for the specific TSN services of the park. In addition, at the protocol stack level, the current QoS is to reserve resources for TSN services through specific identifications. In the future, the TSN service characteristics can be improved to the mobile network QoS system through more refined QoS division, which can ensure the performance of traditional mobile services and TSN services at the same time.

3.4 Space-terrestrial Integrated Network Enhancement

The future space-terrestrial integrated network(STIN) will be a network with underlying terrestrial network and space network as an extension, and a three-dimensional, layered, collaborative, and integrated network. The business mode，technical architecture, standard specification will be unified. Satellites on different orbits (including high-altitude orbits and low-altitude orbits), near-space platforms (such as balloons and UAVs) and nodes on the ground will be organized as a multi-layered network architecture. Network at different layers interwork and collaborate with each other as a globally-covered, opportunistic access, on-demand service, and secure and trustable information network system. The following figure is a reference.

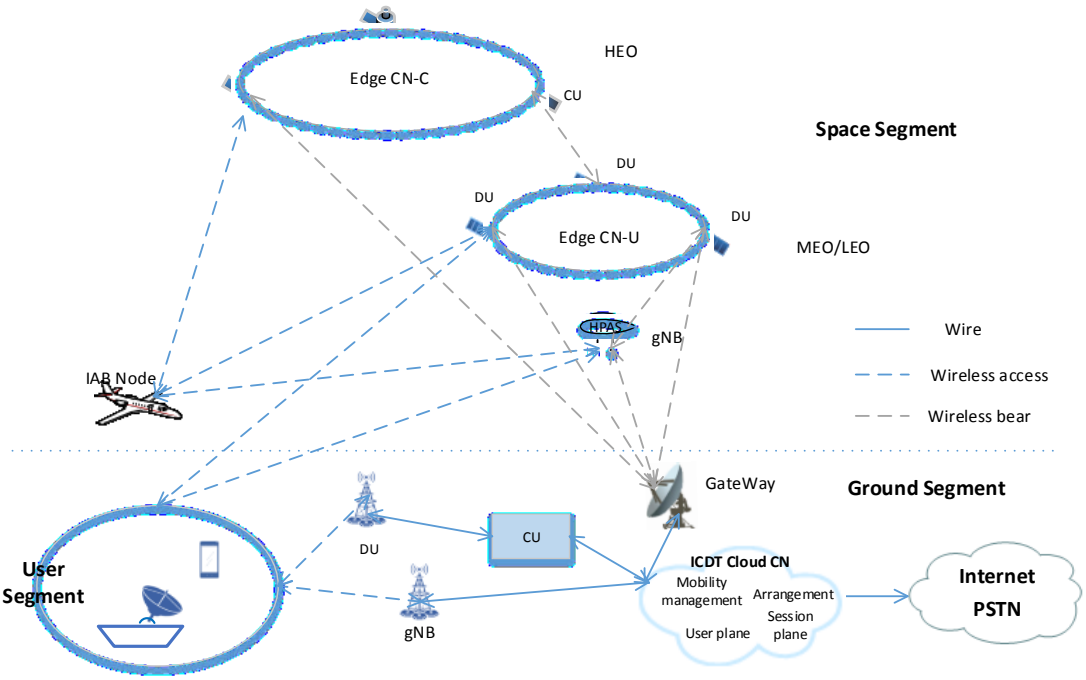

Figure 3.6 A holistic view of space-terrestrial integrated network (STIN)

STIN architecture can be organized based on the physical architecture, logical architecture and implementation architecture.

(1) Physical architecture

For physical architecture, STIN includes three segments namely space segment, ground segment and user segment. The space segment includes space-borne satellites on different orbits and air-borne platforms (such as balloons and UAVs). Ground segment includes satellite gateways on the ground, network OAM system, terrestrial base stations and core network, responsible for the terrestrial communication and the inter-connection to the other network system. User segment includes VSAT terminals, airborne/shipborne/vehicle-borne terminals, and various types of handsets.

(2) Logical architecture

Regarding logical architecture, STIN includes satellite bearer network (S-BN), satellite access network (S-AN), terrestrial access network (T-AN), and terrestrial core network (T-CN).

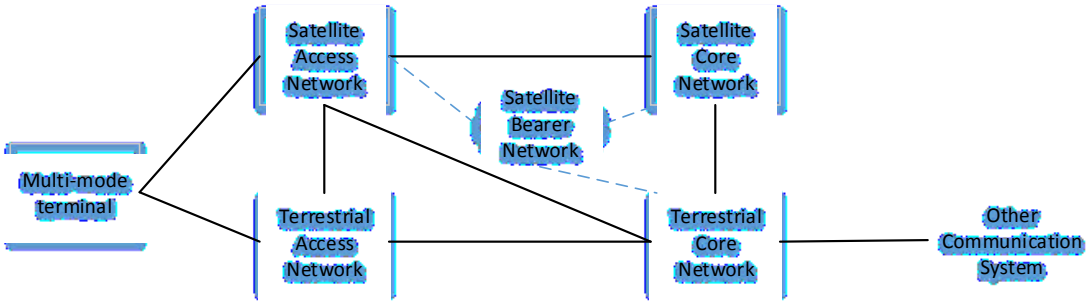

Figure 3.7 Logical architecture of STIN

S-BN includes the space-borne satellites in various orbits, air-borne HAPS and their interconnected network. Every node can have capability of inter-node wide-band wireless communication (via laser or Thz), data transmission relay, routing and

switching, information storage and processing. Via inter-node links various types of user data are forwarded through the space-borne and air-borne network. S-BN nodes also interconnects to terrestrial bearer network through the satellite gateway. Satellite access network includes non-terrestrial network nodes, globally covered and meeting the access requirement of massive users from the ground, sea, air or space. Terrestrial access network refers to the mobile communication network such as 5G/6G. Satellite access network and terrestrial network have the common air interface technology, cooperate with each other, and connects to the core network to realize the interconnection to other network system including internet and PSTN.

(3) Implementation architecture

Regarding implementation architecture, STIN can adopt virtualization technologies such as SDN/NFV. STIN can be orchestrated by mapping or remapping the logical functions to physical entities, to realize the IDCT convergence.

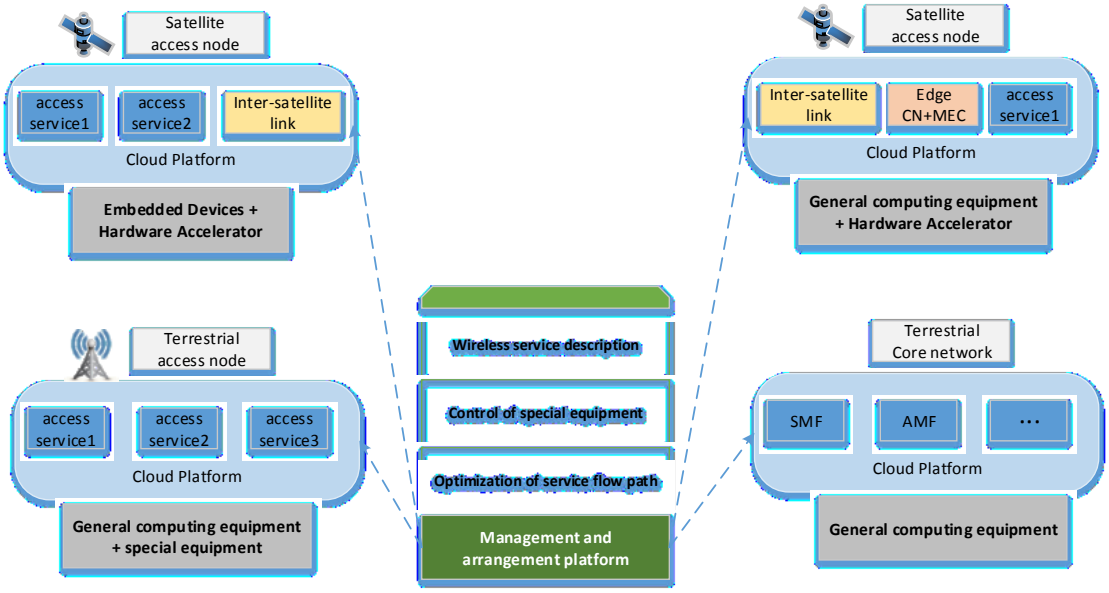

Figure 3.8 Implementation architecture of STIN

The cloud platform includes various resources such as computing/storage equipments, network resources, air-interface resources and energy power. The cloud platform, together with big data and artificial intelligence can be part of the Network as a Service (NaaS) or Network Service Fabric (NSF) to realize the intent-based intelligent network.

For such architecture, STIN can be flexibly configured into the following modes: 1) satellite network is used as a backhaul network for terrestrial network and the terrestrial base station connects to the remote terrestrial core network via the satellite network; 2) part of base station functions like DU is deployed on satellite network and the other part of base station functions like CU is deployed on ground; 3) all base station functions including CU/DU are deployed on satellite; 4) base station functions, some of the core network functions and edge applications are deployed on satellite as the edge cloud.

3.5 Multi-terminal Collaboration

Currently, personal wireless electronic devices are emerging quickly, including: smart watches, smart glasses, wireless headsets, XR helmets, personal health monitoring equipment, home robots, etc. It is foreseeable that in the future, the types of personal devices will be more diverse and smarter. As a result, the necessity of communication among the devices will become higher and higher. The interconnection and collaboration of these devices will greatly empower these smart devices, bringing potential advantages such as improved management and enhanced their communication capabilities.

3.5.1 Scenarios

(1) Media switch seamlessly within multiple devices[3-4]

Since a person has many electronic devices, ideally the user should be able to choose the device he/she wants to watch the video/listen to the audio among all the devices, with simple operation, without interruption of the ongoing media. In other word, the user could switch from one device to anther device and the media continue to play during the switch.

Figure 3.9 Media switch within multiple devices

(2) Direct communication with different RATs[3-5]

There are lots of cases that smart glasses are paired with a smartphone using non-3GPP RAT(e.g. WLAN) for transmitting video information from the smartphone to smart glasses. However, WLAN is based on unlicensed frequency. In some areas, if lots of people use that technology, the quality of service will be bad. If and when the quality of service goes down the service could be switched to a 3GPP RAT direct communication (e.g. sidelink) autonomously and the user could have a better experience. In addition, the opposite could be true in that the direct communications could be congested, and therefore it makes sense that both non-3GPP RAT and 3GPP direct communication could be used together.

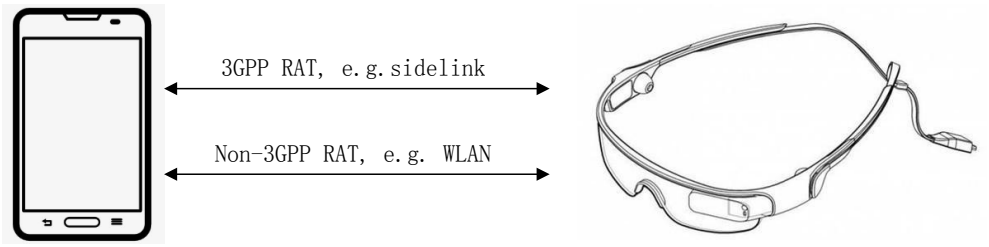

Figure 3.10 Direct communication between devices with different RAT

[3-4] S1-203371 Use case – The Media share within PINs
[3-5] S1- 203372 PINs – Use case – The movie

3.5.2 Architecture and Protocol stack

(1) Inverted dual-connectivity architecture

In 4G and 5G, dual connectivity technology has been introduced, that is, a UE is connected to two base stations at the same time. And the wireless resources of both base stations can be used for the UE's data transmission. According to the available resources of the base stations, the transmission of data between the UE and network can be performed through the primary base station and/or the secondary base station. When the transmission is performed through single base station, the transmission point of the network side can migrate between the primary and secondary base stations seamlessly.

In the scenarios of two terminals collaboration, an inverted dual-connectivity architecture (i.e. the roles of base stations and terminals are reversed) can be adopted to achieve seamless migration of user services between different terminals.

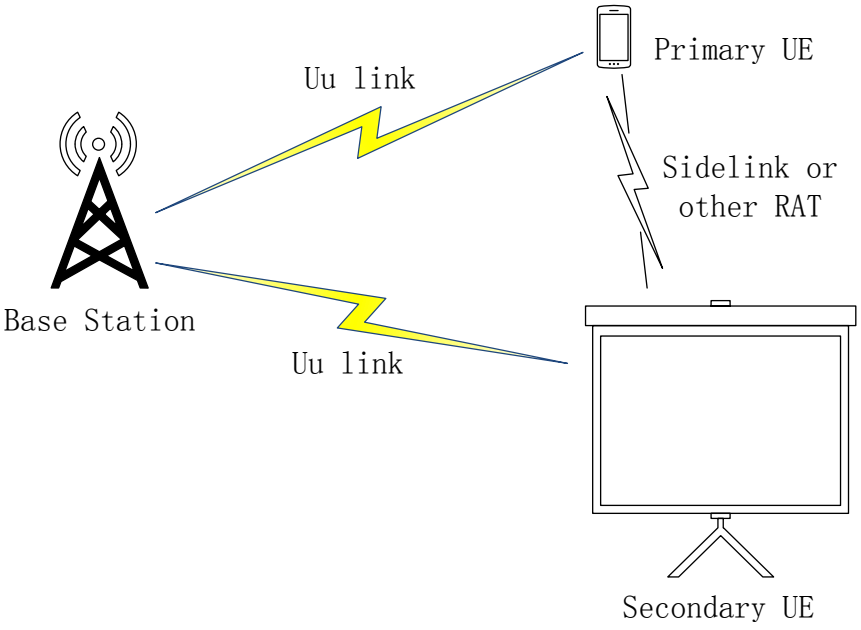

Figure 3.11 Inverted dual-connectivity architecture

In this architecture, two UEs that cooperate with each other are connecting to the same base station. The primary UE maintains a control plane connection with the base station and there is an optional control plane connection between the secondary UE and the base station. When there is no control plane connection between the secondary UE and the base station, the secondary UE can obtain the configuration from the primary UE through air interface.

When the user decides to migrate one service from the source terminal to the target terminal, the target terminal (i.e. secondary UE) for the service migration firstly accesses the base station, and forms an inverted dual-connection with the source terminal (i.e. primary UE). Then, the network migrates the service from the primary UE to the secondary UE. The migration procedure can be similar with the mechanism of seamless migration between the primary and secondary base stations on the network side transmission point of the dual-connection architecture.

This architecture introduces new enhancements to the network-side protocol stack architecture, as shown in the following figure:

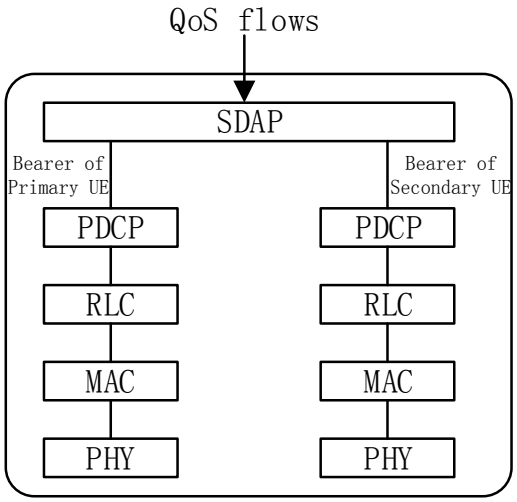

Figure 3.12 protocol stack of network

At the base station, the bearers of the primary and secondary UEs are connected to the same SDAP entity. In this way, data of the same service can be freely migrated between bearers of different UEs.

(2) Multi-RAT aggregation architecture

Direct communication with different RATs can be supported by introducing different RAT aggregation technology. Take sidelink and WLAN as an example: the convergence of sidelink and WLAN means that a backup transmission path based on WLAN should be introduced in the existing 3GPP sidelink work frame. When the WLAN signal quality is good, the data is transmitted based on WLAN, When the WLAN signal quality is poor, the data is transmitted based on the sidelink. In addition to RAT switching, when sidelink and WLAN are aggregated, data can also be transmitted through both sidelink and WLAN simultaneously. Under the control of the network, the split ratio of data on the two RATs can be dynamically adjusted to

achieve high wireless resources efficiency while ensuring user experience.

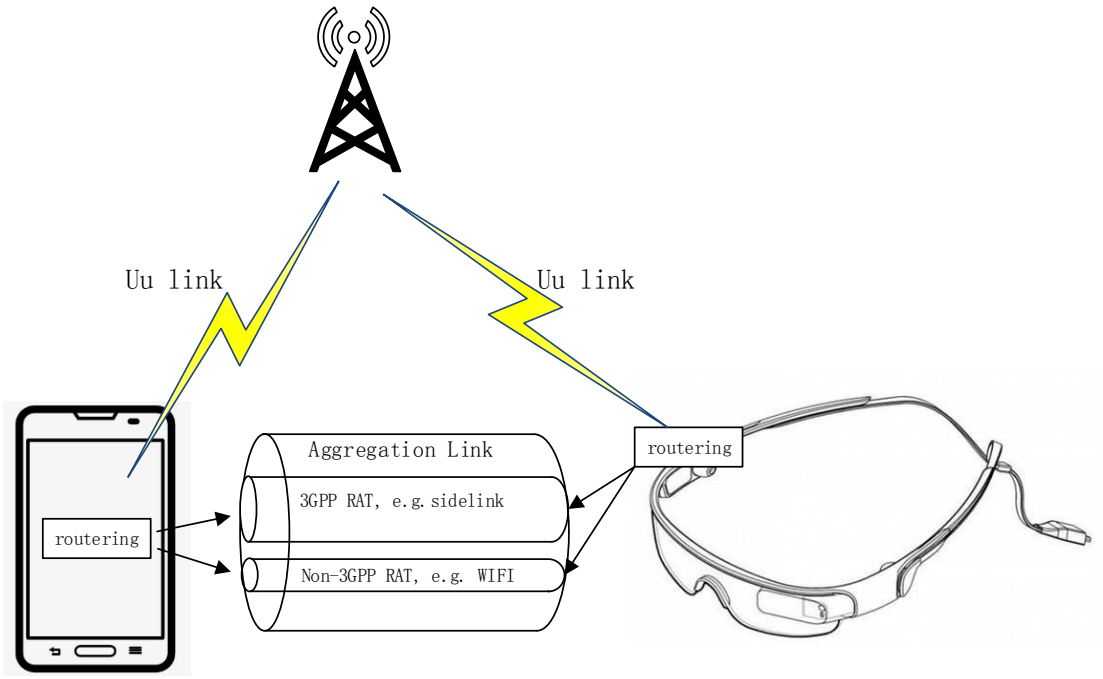

Figure 3.13 communication with multi-RAT aggregation

Under this architecture, the link between UEs uses the following protocol stack architecture (taking the user plane transmission as an example):

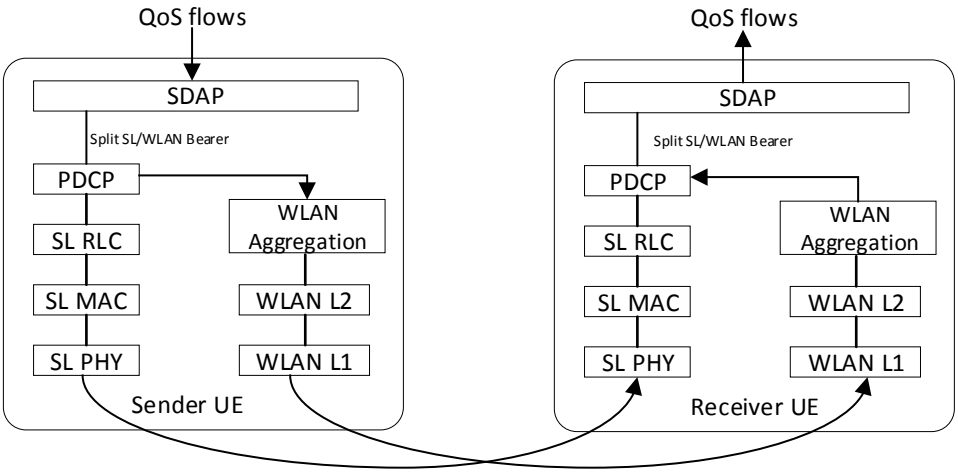

Figure 3.14 SL-WLAN Aggregation

In the PDCP layer of the sender, the routing function is introduced and the data to be sent is routed to the underlying 3GPP RAT protocol stack (such as sidelink RLC) and Non-3GPP RAT protocol stack (such as WLAN) according to the offload strategy configured by the base station. At the receiver, the PDCP layer aggregates the data received from different RATs from the lower layer and delivers them to the top layer.

In the receiver, there may be more than one PDCP entities, each corresponding to

one bearer. In order to ensure that the WLAN L2 at the receiver can deliver the received data packets from lower layer to the correct PDCP entity, a WLAN aggregation layer is introduced between the PDCP and WLAN L2 at both the sender and receiver. The main function of the new layer enables the sender to include the bearer identification information in the packet. According to the identification information, the WLAN aggregation layer at the receiver can route the received packet to correct PDCP entities.

3.6 AI assistant Transmission

Nowadays, AI module has been applied in UEs for more and more purpose, such as image recognition and function recommendation. However, AI module in UE has not been used for communication performance improvement. Hence, the methods to use the information obtained by the terminal AI module (such as the user's daily movement path, etc.) are explored to improve the experience of communication services, network transmission efficiency and other purposes.

3.6.1 **AI assistant transmission**

In the new exciting era of 5G, with support of large bandwidth and low latency, the most expected services have gradually evolved from traditional mobile Internet services (e.g. web browsing) to high-definition multimedia (e.g. 4K video watching) and interactive games based on AR/VR. Accordingly, new services pose new challenges on the 5G networks, e.g. how to avoid network load imbalance, improve spectrum efficiency, and decrease power consumption.

To overcome these challenges, a jointly-optimized transmission strategy, which involves all layers of the protocol stack, such as the Application Layer, the RRC Layer as well as the PHY Layer, can be considered. Currently, the 5G protocol architecture is still based on the traditional layered protocol model, which prevents effective interaction among different layers. For instance, a running APP cannot arrange its data transmission policy (e.g. packet fragmentation), named as distribution policy afterwards, according to the instantaneous network transmission condition (i.e. radio channel condition and network congestion status) due to the lack of information reported from the lower layer. Consequently, real-time matching between the distribution policy for APP data and the network transmission condition can not be achieved. Either the radio resources will not be fully utilized if the distribution policy is conservative (e.g. the data volume of APP data delivered to the lower layer is a bit small), or the radio access network will get stuck in the congestion and more extra power consumption is inevitable if the distribution policy is aggressive (e.g. the data volume of APP data requested from APP server is too large).

The concept of “Intention Network” is introduced, aiming to make the 5G networks to be more intelligent and flexible to satisfy user’s individual service requirements under different network environments at low cost.

3.6.2 Protocol stack architecture

To make the 5G networks become an Intention Network requires some enhancements to the NR control plane protocol stack. As shown in Figure 3.15, the main modification is introducing the Adaption Layer in UE and signaling to indicate the information about transmission condition obtained in the gNB side (i.e. UL radio channel condition and network congestion status).

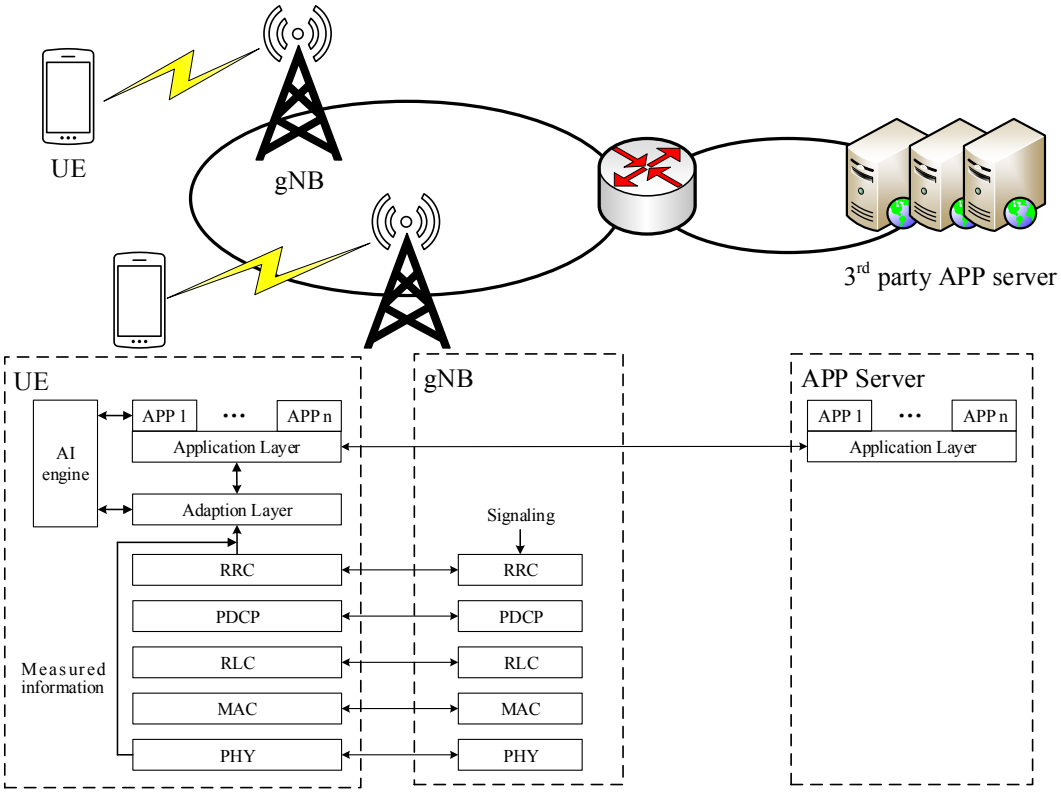

Figure 3.15 Overall architecture of adaptive transmission manager

The adaption layer only exists in the UE side. It is mainly designed to provide the recommended distribution policy to the application layer, based on the information obtained from both the lower layer and the application layer.

Herein, two additional new terms have been derived from the Adaptive Transmission Manager, i.e. the adaption layer and the recommended distribution policy. The following part is intended to clarify the definition of these terms and the corresponding relationships between each other:

• Adaption layer: The adaption layer is the nerve center to recommend a distribution policy for various kinds of services. Based on sufficient information about radio channel condition, network congestion status, and user's individual service requirements, the adaption layer is capable to provide a recommended distribution policy, which is subsequently used by the application layer.

• Recommended distribution policy: A recommended distribution policy is the outcome of the adaption layer and is provided to the application layer. From the perspective of the application layer, the recommended distribution policy can be considered as an important reference to determine the actual distribution policy.

The gNB will transmit necessary transmission information about the UL radio channel condition and the network congestion status to the UE via the abovementioned signaling. Upon the reception of this information, the RRC Layer in UE will deliver it to the Adaption layer. At the same time, the PHY Layer in UE reports the measurement result of DL radio channel (e.g. L1-RSRP) to the Adaption layer. Then the adaption layer collects the information of the service (e.g. service type, user's individual service requirements) via packet inspection and analysis. Based on all the latest information, the adaption layer provides and continuously updates the recommended distribution policy to the application layer. Finally, on referring to the recommended distribution policy, the application layer determines the actual distribution policy, with which the application layer can decide the data volume of APP data should be delivered to the lower layer or requested from the APP server at this moment, and how much APP data packets should be buffered in cache.

In addition, the overall architecture also consists of the AI engine, which plays an important role in information interaction between the adaption layer and the APP. Firstly, the AI engine collects data about the running APPs and the corresponding network transmission condition. It can also analyze and predict user behaviors. After that, the adaption layer can obtain the data statistics from the AI engine, recognizing users' potential service requirements. At last, according to the potential service requirements, the adaption layer is capable to initiate APPs in advance to fetch and store some APP data when it is found out that the network transmission condition is excellent. For example, assuming that a person regularly watches the morning news on the way to office at 8 a.m. on weekdays, the adaption layer will initiate the related APP to download the morning news video at the time ahead of 8 a.m. as long as the network transmission condition is good enough.

3.6.3 Functional description

Adaptive Transmission Manager is envisioned as a platform to efficiently, smoothly, and proactively satisfy users' service requirements. Adaptive Transmission Manager is achieved with the information exchange between the adaption layer and other layers(e.g. application layer, phy layer, RRC layer etc). The adaption layer is consist of 3 modules as illustrated in the following Figure 3.16.

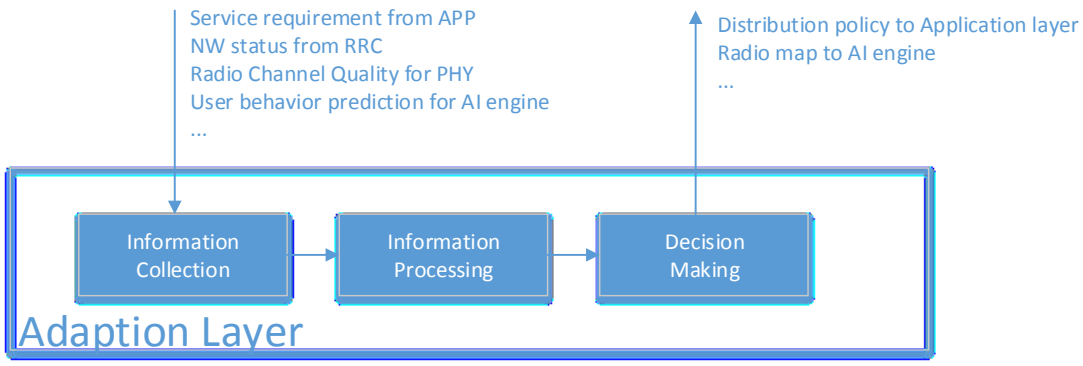

Figure 3.16: Functions of Adaption layer

The first module is information collection, which collects information such as real- time radio channel condition, network congestion status, service requirements, UE location, radio signal strength map and user behavior(e.g. trajectory) predication information.

The second module is information processing. For example, with the service requirements information, the adaption layer can know whether the ongoing service is delay tolerate. With the radio channel condition and network congestion status, the adaption layer can know the optimal data rate the UE can expected at its location currently. With the radio signal strength map and user trajectory predication information, the adaption layer can predict whether user will enter a area which can provide higher data rate in the few minutes.

After information processing, the decision making module will generate distribution policy taking the output of information processing into account. For example, the adaption layer finds the ongoing service is delay tolerate and UE is experiencing low data rate for it is at the edge of the cell coverage, but the UE is expected to enter the center area of the cell very soon. Then the adaption layer may suggest the application layer to halt the current transmission and then continue after high data rate connection is available. In this case, the adaption layer avoids both inefficient radio resource usage and high UE/network power consumption. In another case, the recommended distribution policy may be different. If the ongoing service is delay sensitive, the adaption layer may suggest the application layer the reduce the codec rate when the available date rate is limited.

It can be found that the adaption layer helps the application layer to determine a comprehensive distribution policy based on the information about both network transmission conditions from the lower layer and the user's individual service requirements. Due to the time-varying characteristics of radio channel condition, network congestion status, the recommended distribution policy is continuously updated by the adaption layer accordingly.

In legacy protocol stack, information exchange between the application layer and the radio access layer is very limited, which prevents further optimization on the radio resource efficiency and user experience. The Adaptive Transmission Manager provides a way to solve this problem. What's more, the Adaptive Transmission Manager also takes the prediction on user behaviors into account. Based on the history statistics of user behavior obtained by local AI engine, UE can predict the potential service requirement from user and the potential change of radio environments. Therefore, Adaptive Transmission Manager makes it possible to generate distribution policy in a proactive way, which is a very promising direction for further research.

3.7 HTC Support

3.7.1 Scenario

Before the new technology of HTC (Holographic-Type Communications) appeared, the application scenarios of HTC (Holographic-Type Communications) have already shown up. In science fiction books and some films and television programs, HTC technology has brought all-round impacts on social structure, production and life. HTC realizes the reconstruction of real objects through three-dimensional holographic images and perception technology, and then establishes an interactive feedback information loop between the virtual world, the real world and users using the real-time capture technology and transmission technology, as well as high-resolution imaging technology. Applying HTC to the Next-generation communication network can enable users to enjoy an immersive experience completely, and realize the deep integration of virtual and real scene.

HTC technology can be widely used in many fields such as culture and entertainment, health care, education, military, and social economy in the future. It will become the main form of dynamic three-dimensional presentation of people, objects and the surrounding environment through its natural and realistic visual restoration, which greatly meets the needs of humans for multi-sensory integration between human beings and their environments. For example, applied to entertainment scenario, including multimedia services, remote meetings, home communication and other scenarios, HTC will truly realize the three-dimensional complete construction and presentation, as well as real-time interaction and communication, etc. applied to industrial scenario, including industrial modeling, exhibition design, advertising and other scenes, HTC will bring people a brand new upgraded experience, which combines with multi-dimensional sensory information; applied to the public services scenarios, including health care and telemedicine, etc. HTC makes holographic microscopy, telemedicine surgery and “digital human” e-health possible, promotes the development of biomedicine, and protects human health.

In general, HTC technology includes the following major aspects:

- Data collection and processing, involving three-dimensional data modeling, intelligent perception and other technologies;

- Data transmission, which is the information transmission and processing process directly participated by the network, sending a huge data stream from the remote end to the presentation end;

- Data presentation, namely three-dimensional display, involving multiple emerging technologies such as data reconstruction, holographic data presentation and voice reconstruction, etc.

Related to the protocol stack is mainly the part of the HTC data transmitted in the network. The basic requirements of HTC include:

- Ultra-high bandwidth: 1Gbps~1Tbps
- Ultra-low latency: 1~5ms
- Support for concurrent flows. Depending on point cloud and image array dimensions, on the order of 1000 concurrent flows may need to be supported.
- User experience transmission rate: 10~20Gbps
- Ultra-low miss rate: approximately 0

3.7.2 Impacts on Architecture and Protocols

The influences of HTC on the architecture and protocols are reflected in the following aspects:

- HTC services have a multi-angle and all-round three-dimensional presentation effect, and the multi-dimensional presentation poses new challenges and thinking on the design modeling of the presentation side;
- The requirements of ultra-large data traffic and ultra-high real-time pose challenges on the design of transmission networks;
- The collaboration and linkage of multi-flows pose new ideas for protocol stack design;

3.7.3 Design of Architecture and Protocols

1. Architecture Design

In order to achieve the transmission requirements of HTC services, etc. a new RAN architecture model that uses anchor UE to trigger multi-network node cooperative transmission is proposed. When the user's anchor UE has new service transmission requirements such as HTC, the network side triggers multi-network node to transmit services cooperatively, according to the location information of the anchor UE. Particularly, the network node devices that can be cooperative transmitted include network-side nodes that can perform data collection and presentation, terminal devices of other users that can be used for coordinated transmission, and sensor devices that have the functions of collection and presentation, etc. the specific architecture is shown in the figure below.

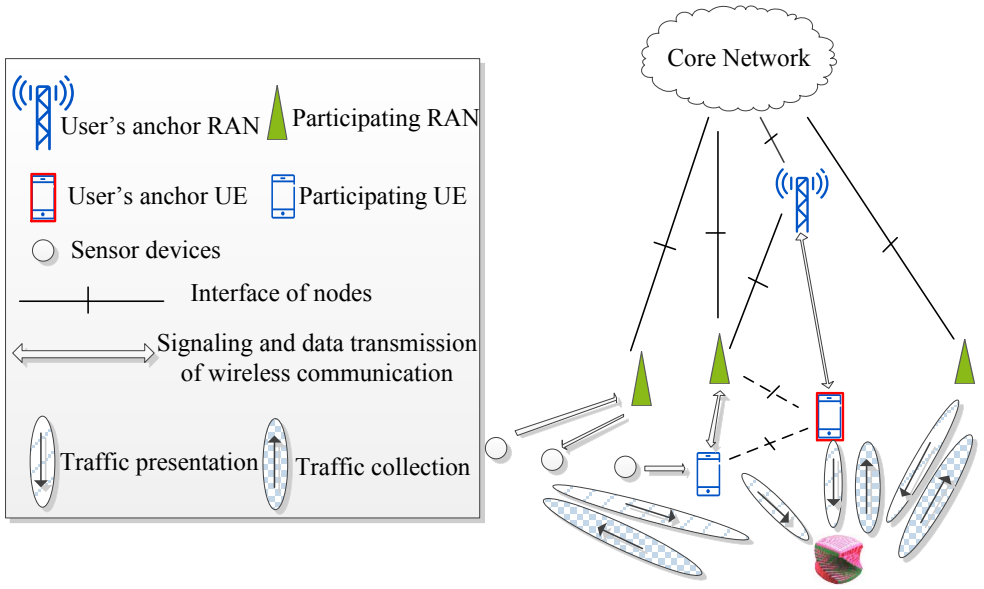

Figure 3.17 Overall architecture of HTC support

Compared with the traditional mobile communication network architecture, the multi-network nodes in the above architecture cooperatively transmit user services, increase the amount of data that users can transmit, increase service transmission rate and reliability, reduce data transmission delay, and enhance user experience. And meet the transmission needs of user services. For new services such as HTC, the user data that can be collected and presented is more detailed and comprehensive. In view of the multi-angle and all-round three-dimensional presentation effect of new services such as HTC, it is difficult to meet the multi-dimensional presentation with a single terminal device, limited by the conditional of terminal hardware and capabilities of transmission and processing. Therefore, in the architecture model, the anchor UE can transmit and present with other network nodes which have the collection and presentation functions cooperatively, to perform multi-angel and omnidirectional three-dimensional presentation of new services such as HTC.

In view of the network transmission requirements such as ultra-large data traffic and ultra-high real-time of the HTC and other new services, it poses great challenges to both the user equipment and the network equipment. For example, the transmission of super hi-vision and real-time HTC services with human size will require a transmission rate of at least Tbps level. In consideration of the energy consumption, transmission rate limitation and other issues of user equipment, a signal user equipment alone cannot meet the ultra-large data transmission requirements of new services such as HTC, and thus cannot meet the service transmission characteristics of real-time and reliability. Therefore, in the architecture model, the anchor terminal can trigger multiple network node devices to coordinate the transmission of user service such as HTC. Each related node device can receive or send part of the data transmitted by the user which is more or more data streams of HTC services. And the cooperative transmission services data between devices is performed by the network side for unified parameter configuration, coordination and scheduling, etc. Through

coordinated data transmission of multiple network nodes and related devices, the transmission performance such as ultra-high transmission rate, ultra-low latency, and ultra-high reliability of HTC services is guaranteed.

2. Protocol Design

Multi-stream linkage architecture requires flexible design and deployment of protocol stacks of different participating nodes. According to the different requirements of data collection, transmission and presentation, different nodes can deploy different protocol stacks, and the protocol stacks between related nodes need to be appropriately associated and deployed.

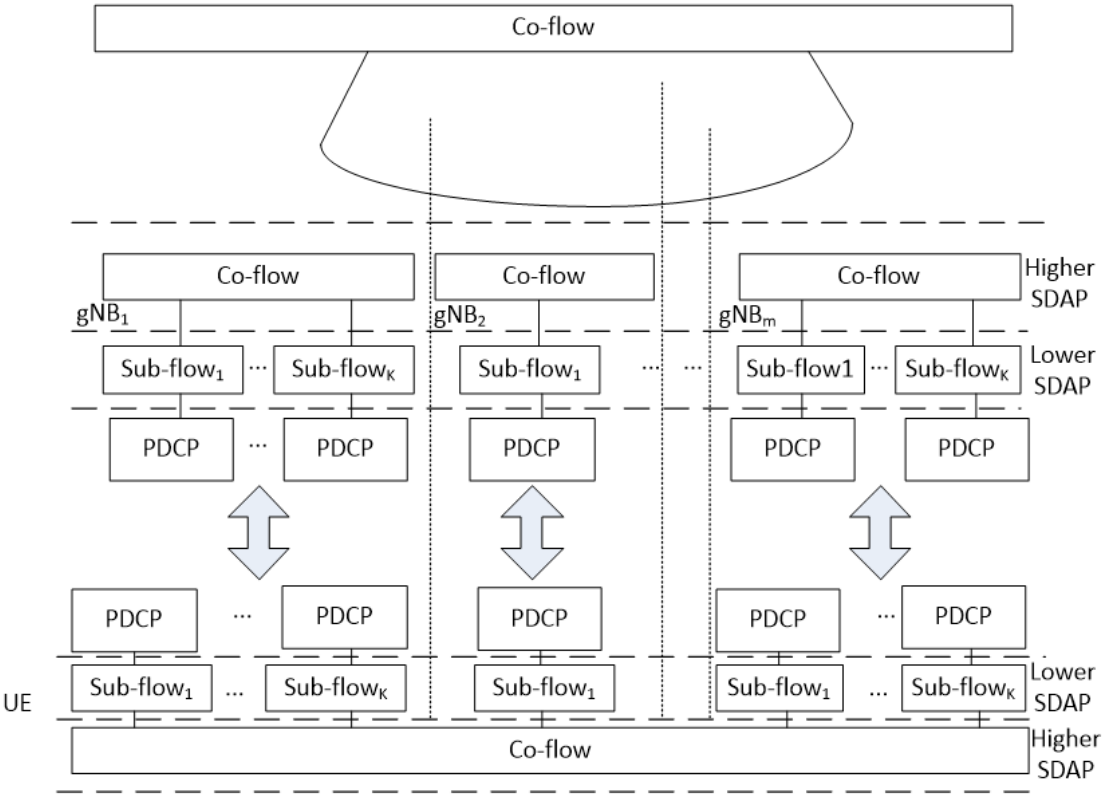

Figure 3.18 Protocol stacks of HTC support

The concept of co-flow is introduced in HTC, multiple flows of co-flows are related to each other, and they can have different QoS requirements. From the perspective of holographic projection and sampling, RAN-side data flows can come from or sent to different peer nodes. The figure above is a schematic diagram of UE protocol stack which receiving different sub-flow that belong to the same co-flow, and this sub-flows can come from different gNBs. The sub-flow under the co-flow needs to realize linkage and QoS guarantee, and the network and the terminal can configure the sub-flow under the co-flow connection.

3.8 Separation of user and UEs

3.8.1 Scenario

With the introduction of massive smart terminal devices, people's lives have become more flexible and intelligent. In the communication system, a user can have multiple smart terminal devices in different locations, such as computers, mobile phones, entertainment wearable devices, smart home devices, etc. in the home environment, tablets, notebooks, smart displays and operable smart devices, etc. in the office environment, as well as smart devices in the private car environment such as vehicles and displays. These separate terminal devices served for a user jointly, and users can activate one or more terminal devices for use with different needs.

At present, Internet supports this function, For example, a user can log in to watch a video on the video APP of one device, and then when using another device, the user can enjoy continuous video logging in to the same account. But this does not realize the continuity of service transmission in the mobile communication system. And the continuity of service at the application layer has a large latency and complicated user operations, as well as limited applications.

The future mobile network needs to achieve user-centric services continuity, so that users can enjoy the continuous and even non-destructive experiences, such as official business, remote meetings and entertainment etc., on separate terminal devices when they move at home, walk out of the house, drive a vehicle or go to the company office, thereby improving the life quality and work efficiency of users.

3.8.2 Impacts on Architecture and Protocols

Generally speaking, User and UE are the same concept in wireless communication networks. The scenario where multiple UEs serve one or a group of users will naturally trigger the idea of separating users from terminals. The continuity of service data of multiple UEs serving the same User can be realized on the RAN side, which can improve the user experience and greatly increase the application range of various services, and the users can receive services from multiple terminals seamlessly.

Under the User/UE split architecture, the relationship between the user and UE is mainly that the user can trigger the associated UE to transmit air interface data and signaling. And according to different application scenarios, there can be signaling and data interaction, or not.

Case 1: No user plane data transmission between User and UE. For example, user (human) attends a video meeting by deployed fixed or mobile UEs on the way when it is moving. The received data present (video) on the screens of activated UEs directly and data from user to NW are collected by the activated UEs. In this case, no user plane data transmission is needed between User and UEs but User needs to activate necessary UE(s) when needed.

Depends on the interaction between User and UE:

- If User activates UE(s) by non-3GPP ways, such as face recognition, fingerprint, etc, there is no impact on protocols of wireless network. The impacts on wireless network are configuration of protocol parameters, the mechanism of UEs activation/deactivation, and etc.

- If User activates UE(s) by 3GPP signaling, it has impacts on protocol design.

Case 2：User plane data are transmitted between User and UE. For example, User is robust or other kinds of data collection/ control system. Besides activating the UEs on the way, user need to acquire data or sent command signaling to the UEs. It has protocol impacts.

3.8.3 Design of Architecture and Protocols

1. Architecture Design

To satisfy the service continuity of User/UE split, new network architecture can be considered. In this architecture, a User can access to different UEs in TDM way. The UEs associated with a user can maintain service continuity in RAN level. Another use case is a User accesses to multiple UEs and each UE provides part of service to the User. The third use case is a group of Users with same or similar characteristics can access to one or a group of UEs to get required service, such as, surgery staff in one operation and different division of labor in the manufacturing of a project, etc.

For example, UEs associated to a User can be deployed in different location. The User activates specific UE(s) for required services by user account, body characteristics, dynamic or semi-static password, and etc. Data transmission for the User is performed in the activated UE(s). With mobility of the User, different UEs can be activated. The service continuity among the old and new activated UEs should be considered to improve the user experience.

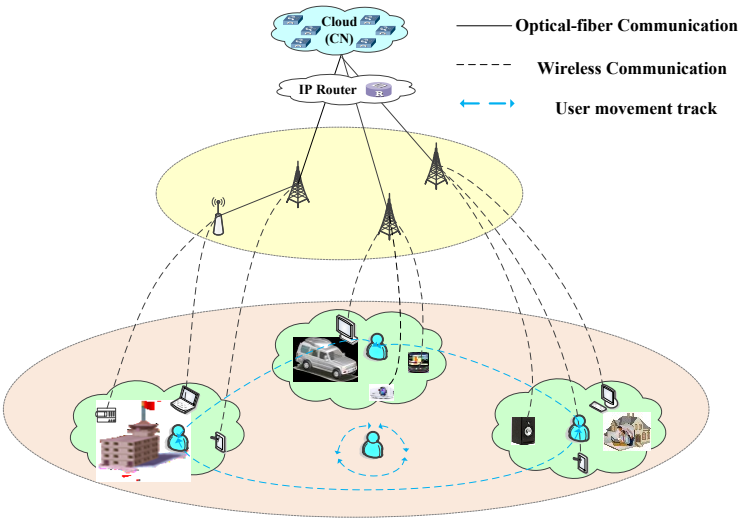

Figure 3.19 Overall architecture of separation of user and UEs

In this architecture, initial attachment between User and UEs is setup by user account, fingerprint/face recognition/voice identification, password, or other authorized ways. After initial attachment, User can activate associated UE(s) for needed services. The activated UEs will access to RAN to get data transmission for the User.

During mobility, a User can access to different UEs. When User moves from one UE to another while data transmission is ongoing, service continuity is an important issue. Service continuity can be considered from below aspects:

- Forward data between UEs directly;
- Forward data via Networks.

In the User/ UE split architecture, User is associated to multiple UEs and the UEs can be joint-configured and coordinated for better service to one User.

In the case that users in one user group get service from same UE or UE group, User Group ID is needed. The UEs can perform initial attachment to one user with the User Group ID and then serve all users with the same User Group ID.

2. Protocol Design

For the cases in clause 3.8.2, there is no data interaction in case 1, but if user activate and change UE(s) by 3GPP signaling, control plane data transmission is needed. The new RRC signaling includes UE activation/deactivation, user context, verification information, and so on. The new RRC should include the information and procedures for initial attachment.

An example of new RRC protocol is shown in below figure.

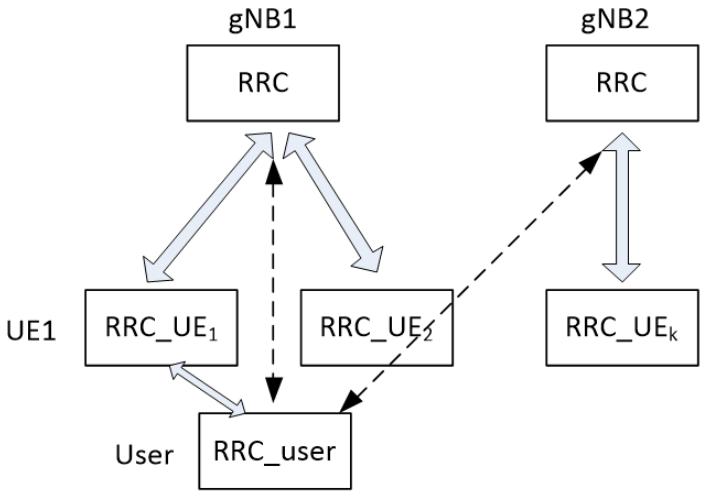

Figure 3.20 Protocol stacks of separation of user and UEs

When user plane data is needed between User and UE as case 2 in clause 3.8.2, the basic protocol types are same to 5G but it can be simplified according to the data types.

4 *Summary*

The Protocol Stack White Paper 2.0 analyzes the main problems faced by existing networks and important scenarios for future network deployment to explore better protocol stack architecture and functions. One of the main issues discussed in the Protocol Stack White Paper 2.0 is how to ensure high-reliability data transmission and the ultra-low latency requirements of a single service while meeting lower transmission resource costs. The multiple solutions proposed by the Protocol Stack White Paper 2.0 include simultaneous transmission of multiple connections, multiple transmissions on a single link, and a solution from the MAC and PHY protocol layers.

This white paper further studies the protocol stack architecture and protocol layer functions, and analyzes some problems in the current network, such as the deficiencies of CU/DU in cloud, the deficiencies of edge NWDAF, the problems faced by core network slicing, and the problems faced by access network slicing, etc. The white paper analyzes the next-generation protocol stack from three directions of service, service-based, component-based and intelligence-based. In addition, this white paper analyzes the difficulties faced by the existing networks one by one, and provides some solutions as well as exploration direction, based on the current scenes and technologies that 3GPP pays attention to, such as NPN network enhancement, intelligent enabling network enhancement, TSN network enhancement, etc.,.

The consumers' individualized demand for communication services, and vertical industries' demand for diversified communication scenarios, all place increasingly stringent requirements on the standards and technologies of the communication industry. The exploration of better protocol stack solutions and general protocol stack architecture is the eternal driving force for the continuous development of the communications industry and technology.

We hope this white paper inspire more communication practitioners to work together and let’s create a more colorful communication world.

5 *Reference*

[1] 5G White Paper：Next-Generation Protocol Stack

[2] 5G White Paper：Next-Generation Protocol Stack 2.0

6 *Abbreviation*

6G	The Sixth Generation Mobile Communications
CU	Central Unit
DU	Distributed Unit
NWDAF	Network Data Analytics Function
SBA	Service-based Architecture
CN	Core Network
TSN	Time Sensitive Network
NTN	Non-Terrestrial Network
NPN	Non-Public Network
AI	Artificial Intelligence
RLC	Radio Link Control
RRC	Radio Resource Control
UDP	User Datagram Protocol
UE	User Equipment
UM	Unacknowledged Mode
AM	Acknowledged Mode
ARQ	Automatic Repeat request
PDU	Protocol Data Unite
UE	User Equipment
UM	Unacknowledged Mode
BWP	Bandwidth Part
MCG	Master Cell Group
SCG	Secondary Cell Group
SR	Scheduling Request
AM	Acknowledged Mode
DRB	Data Radio Bearer carrying user plane data
gNB	NR Node B
IP	Internet Protocol
MAC	Medium Access Control
PDCP	Packet Data Convergence Protocol
RB	Radio Bearer
HTC	Holographic-Type Communications

7 *Acknowledgement*

Grateful thanks to the following contributors for their wonderful work on this whitepaper:

Whitepaper:

Editors：Chih-Lin I, Junshuai Sun, Huimin Zhang

Contributors：

China Mobile: Guangyi Liu, Qixing Wang, Jing Jin, Na Li, Yingying Wang, Yun Zhao, Juan Deng, Quan Zhao, Min Yan, Xuan Liu, Xin Sun, Qingbi Zheng

ZTE: Feng Xie, Kaibo Tian

Datang Mobile: Li Chen, Yan Wang

vivo: Yitao Mo, Yanxia Zhang, Jiamin Liu, Wei Bao

China Unicom: Shan Liu, Rong Huang

China Telcom: Xiaoyu Qiao, Jing Wang, Jiaxiang Liu

FuTURE FORUM is committed to cutting edge technologies study and applications. Controversies on some technical road-maps and methodologies may arise from time to time. FuTURE FORUM encourages open discussion and exchange of ideas at all levels. The White Paper released by FuTURE FORUM represents the opinions which were agreed upon by all participating organizations and were supported by the majority of FuTURE FORUM members. The opinions contained in the White Paper does not necessarily represent a unanimous agreement of all FuTURE FORUM members. FuTURE FORUM welcomes all experts and scholars' active participation in follow-on working group meetings and workshops. we also highly appreciate your valuable contribution to the FuTURE White Paper series.

未来移动通信论坛

FUTURE MOBILE COMMUNICATION FORUM